有趣的近边界数据：模型所有权推断 Interesting Near-boundary Data: DNN Model Ownership Inference

孙哲

2023 年 2 月 2 日

1 Abstract

深度神经网络（DNN）的训练代价昂贵是导致模型知识产权保护问题逐渐被重视的原因。近些年来，模型盗窃时常出现，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指纹用于验证模型所有权。然而，歧义性声明等攻击手段被用于破解模型水印和指纹，这对模型所有权验证工作造成了挑战。因此，本文提出了一种有趣的近边界数据作为获得模型所有权的证据，并创新性地提出了推断模型所有权而不是验证模型所有权。本文提出采用对抗性样本生成算法和生成对抗网络构造私有化的近边界数据，我们主要的观察结果是近边界数据在源模型和其衍生的盗窃模型中均表现出靠近分类边界的结果。在这一项工作的最后，我们设计了大量的实验验证提出方法的有效性，注意我们的方法不要求训练数据私有。实验结果证明近边界数据在源模型和盗窃模型具有相同的近边界特性，并且可以利用该特性推断模型所有权，使用者可以使用近边界数据作为模型知识产权保护的技术支持。

The high training cost of deep neural network (DNN) is the reason why the model intellectual property (IP) protection is gradually paid more attention. Recently model stolen has frequently occurred, and many researchers have been inspired by digital media watermarking to design model watermarks and fingerprints to verify model ownership. However, attacks such as ambiguous claims are used to crack model watermarks and fingerprints, which pose a challenge to model ownership verification. Therefore, this paper presents an interesting near-boundary data as evidence for acquiring model ownership, and innovatively proposes inferring model ownership rather than verifying model ownership. In this paper, we propose to employ adversarial samples generation algorithms and generative adversarial networks (GAN) to construct privatized near-boundary data. Our main observation is that near-boundary data exhibits results close to the classification boundary in both the source model and its derived stolen model. At the end of this work, we design extensive experiments to verify the effectiveness of the proposed method, noting that our method does not require training data to be private. The experimental results show that the near-boundary data has the same near-boundary characteristics in the source model and the stolen model and the model ownership can be inferred by using this characteristic, and users can use the near-boundary data as a technical support for model IP protection.

2 Introduction

作为一种数字产品，深度神经网络 (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计算资源 [Co 17]。然而，大多数模型经常被暴露给公众以提供机器翻译 [] 或图像识别 [Co 25] 等服务，这使得攻击者有机会通过暴露的接口窃取模型 [Chandrasekaran et al. 2020]。在这种情况下，盗窃者仅从对源模型的 API 访问中派生出代理模型 [Jia et al. 2021]。因此提出了一个所有权解决问题：所有者如何证

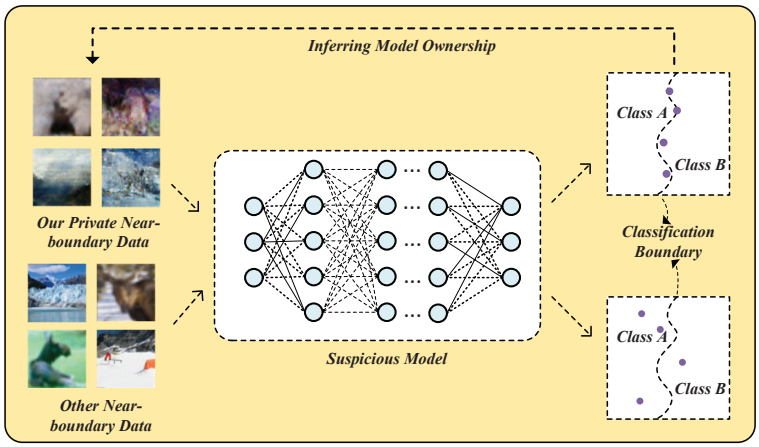


图 1: Inferring model ownership with near-boundary data.

明一个可疑模型窃取了他们的知识产权。具体地说，我们的目标是确定潜在的被盗模型是否是从所有者的模型或数据集派生出来的。

盗窃者可能以多种方式派生和窃取受害者的知识产权 [Fang et al. 2019]。一种突出的方式是模型蒸馏 [Hinton, Vinyals, and Dean 2014]，其中盗窃者利用对模型的预测结果的访问 (例如通过 API) 来以比开发模型所致的成本更低的成本来复制模型。此外，盗窃者可以完全访问受害者模型 [zongshu1_2and 4]，但不能访问数据集，当受害者希望将他们的工作开源用于学术目的，盗窃者可能对受害者模型微调或模型剪枝。

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 [zhiwen 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。Motivated by this，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。An illustration of the idea is shown in Figure 1.

Contributions. The contributions of this paper are: 1) 我们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 (Section); 2) 我们提出了利用对抗性样本构造近边界数据以抵御模型窃取 (Section 3); 3 我们设计了基于 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。我们还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. 3) 我们进行了广泛的实验在 ResNet18 上，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。(Section 4)。

3 THEORETICAL MOTIVATION

3.1 所有权验证局限

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（robust）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 [DeepAuth 有引用] 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 additional 水印和指纹混淆所有权验证。具体来说 Secifically，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 DNN IP 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，robust watermark 可能不一定会验证所有权，除非水印方案是不可逆的 [综述 1 的 46]。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 Dataset inference[] 提出的 ownership resolution，我们将在下一节中具体讨论。

3.2 利用数据推断所有权

Dataset inference 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 ownership resolution。

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 ownership resolution，其中 resolution 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。We can observe that 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。

数据集推理的局限性：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[水印 8] 等人验证了此限制，结果表明该方法产生的结果值得怀疑 questionable。

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。

3.3 Threat model

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的目标，我们定义了成功推断模型所有权的要求：

Fidelity. 推断模型所有权的方法不应影响模型的性能，模型的最大可接受测试精度下降不超过 5pp。

Effectiveness. 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。

Robustness. 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。

Invisibility. 敌手无法获得私有的近边界数据，也无法观察到近边界数据的部署在视觉上。

Efficiency. 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。

Problem Definition

我们定义了一个 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 c-class DNN classifier，分类器输出层为 softmax 层或其他决策层，决策函数 $g_j(x)$ 表示样本 x 被分到 j-th class 的概率，其中 $j = 1, 2, .., c$。$Z_1, Z_2, ..., Z_c$ 表示源模型分类器的全部决策函数的输出，其结果可作为分类边界的依据被我们使用，因此

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

注意，样本 x 的标签 y 被推断为拥有最大概率的类别，例如 $y = argmax_j g_j(x) = argmax_j Z_j(x)$。

Definition 1 (Classification Boundary). 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。在过往的工作中 [指纹 0] 分类边界被描述为如果一个样本数据具有两个相等的最大概率（or logit），则该样本数据位于分类边界上。因此，我们可以使用一组位于分类边界上的数据点抽象表示分类边界。注意，分类边界是客观存在的，和是否能够找到这样的数据点们并不相关。

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，我们将在 Section 4.1 中讨论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。

4 近边界算法 for Ownership

基于 Section 3 中的讨论，我们提出构造近边界数据从而推断模型所有权，而不是验证模型所有权。具体而言，如图2所示，我们的方法包括三个主要阶段，包括 (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，以及 (3) 加入近边界数据微调源模型。其技术细节见以下小节。

4.1 近边界对抗性样本

在本节中，我们会将描述如何构造近边界数据的第一步，生成靠近分类边界的基本数据点。在讨论具体操作之前，我们首先总结近边界数据的概念并给出定义。

Definition 2 (Near-boundary Data). 给定一个样本 x，一个阈值 τ，如果样本 x 满足 $|g_i(x) - g_j(x)| \leq \tau$，其中 $i \neq j$ and $\min(g_i(x), g_j(x)) \geq \max_{k \neq i,j} g_k(x)$，则样本 x 被称为近边界数据。

Section 5 和附录中，我们通过大量的实验证明了近边界数据在大多数的模型窃取技术中其近边界特征被保留。因此，近边界数据可以作为推断所有权的依据被使用。尽管近边界数据的特征在模型 IP 保护中表现出显著的效果，由于自然的近边界数据在样本空间中的占比较低甚至可以被忽略，因此如何得到近边

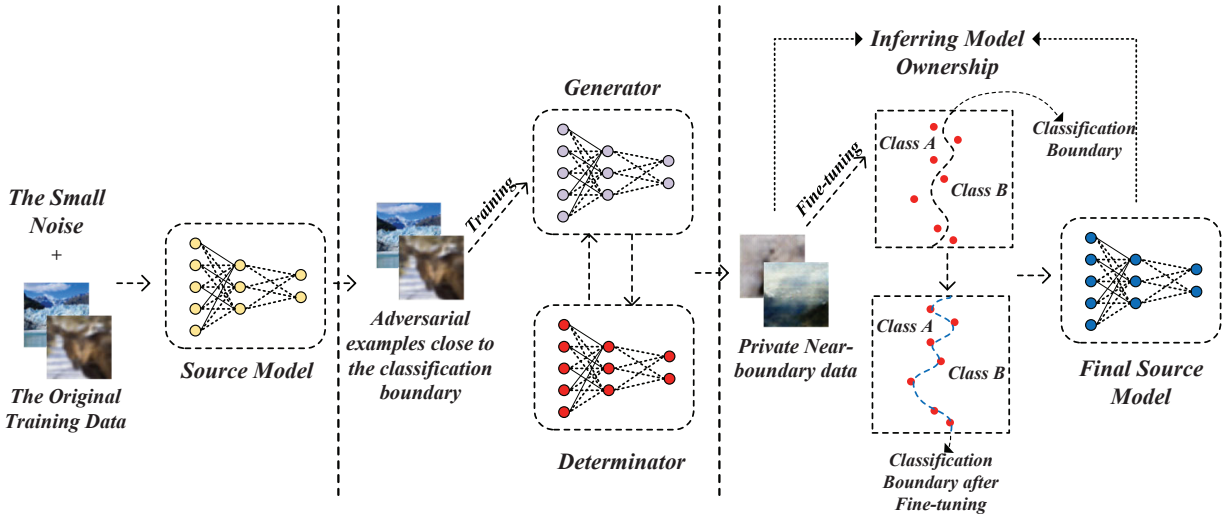


图 2: Construct private near-boundary data to infer model ownership.

界样本仍然很困难。然而，根据最近的一些研究 [指纹 0] 我们知道对抗性样本常被用于确定分类器的分类边界。对抗性样本本质上是通过向样本中添加精心设计的噪声，使一个样本跨越分类器的分类边界。具体而言，对抗性样本有两个分类：原分类和目标分类。原分类是该样本原始的分类结果，而目标分类则是添加噪声后的分类。对抗性样本对于分类边界的跨越体现在在视觉中仍是原分类但分类器的结果却是目标分类。本文认为该特征可以帮助从对抗性样本中得到较多的近边界数据。因此，我们测试了几种常用的生成对抗性样本的方法 [指纹 0 3，8，21]，以帮助我们构建近边界数据。我们专注于有针对性的方法，因为希望可以得到任意分类边界附近的数据，这对我们后续的思路有极大的帮助。

Fast Gradient Sign Method (FGSM).FGSM[zhiwen0 8] 是最经典的构建对抗性样本的方法，只需对初始样本添加 a noise bound ϵ, 如下式2中 FGSM 返回一个对抗性样本 $x^{'}$:

$$x^{'} = clip(x - \epsilon \cdot sign(\bigtriangledown_x J(C, y^*; x))) \tag{2}$$

where $clip$ is the function to the function to project the adversarial example back to the feasible data domain, $\bigtriangledown_x$ is the gradient, J represents the loss function of classifier C, and y^* is the target label.

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[21] 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，这和我们指定目标分类结果有关。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[zhiwen0 3] 能够生成大量在分类边界附近的样本，具体的测试结果被放在附录中。

Carlini and Wagner' s methods (CW).CW 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 and CW-L_∞, 不同的变体使用不同的方法 measure the magnitude of the noise，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，具体操作将在下一小节中给出。

Dataset	Acc. before Fine-tuning	Acc. after Fine-tuning
CIFAR-10	0.886	0.873
Heritage	0.879	0.856
Intel_images	0.794	0.786

表 1: The influence of fine-tuning the target classification boundary on the source model.

4.2 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 Section 4.1 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同样是近边界数据对于源模型。因此，在本文中我们设计了一种基于 DCGAN[] 的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-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 fake 数据，判定器判断生成器的结果。DCGAN 的目标函数如式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。

$$\begin{aligned} &min_G max_D V(D,G) = min_G max_D E_{x\sim P_{data}(x)}[logD(x)]+ \\ &E_{T\sim P_T(T)}[log(1-D(G(T)))] \end{aligned} \tag{3}$$

其中 T 是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。噪声 T 的维数至少要达到数据流形的内在维数，才能够生成我们需要的近边界数据。注意这里的优化过程是一个交替过程。

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，考虑到在两者效率较高的情况下，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，如式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，$g_t(\cdot)$ and $g_s(\cdot)$ 分别表示目标分类概率和源分类概率 (or logit)，$Loss_{FT}$ 本质是希望近边界数据更靠近目标分类边界。$Loss_{SM}$ 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。

$$Loss_{FT} = \frac{1}{n}\sum_{i=1}^{n}(g_t(x_i^{'}) - g_s(x_i^{'}))^2 \tag{4}$$

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，由于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1所示，微调前后源模型的精度差不超过 $5pp$。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附录中找到。

4.3 Inference Ownership with Near-boundary data

在本文中，我们通过生成对抗性样本，私有化近边界数据和微调目标分类边界三个阶段构建了私有化的近边界数据用以推断模型所有权。如 Section 3.2 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。

Hypothesis Test. 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-value，以及 the effect size $\Delta\mu = \mu_S - \mu$，$\Delta\mu$ 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。

5 Experiments

5.1 Experiments Settings

本文选择 CIFAR-10[脚注],Heritage[] 和 Intel_image[] 等开源数据集作为评估的依据,并且选择 ResNet18[] 作为评估的源模型，VGG11[] 作为无关模型。本文所使用的模型均在开源的预训练模型上进行训练。注意本文的全部代码开源 []。

被盗模型 Selection. 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。注意由于我们提出了微调目标分类边界，因此还设置了多种目标分类边界及微调后的源模型。更多的实验设置包括模型参数、超参数等可以在附录中找到。

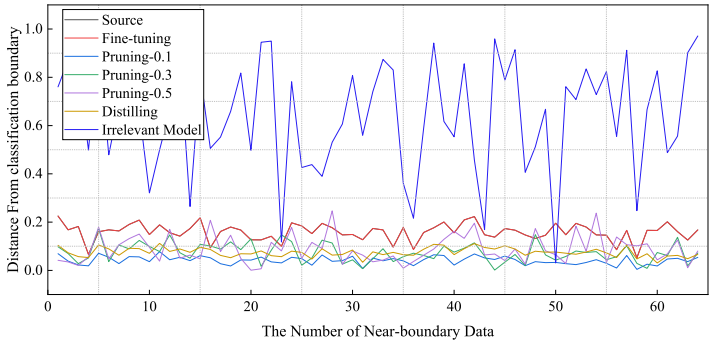
本文遵循 [所有权验证] 中使用 the effect size $\Delta\mu$ 和 p-value 对实验结果进行评估。具体实验中每次采样 10 个数据的结果计算 $\Delta\mu$ 和 p-value，通常来说，$\Delta\mu$ 越大越好，p-value 越小越好，表明我们的方法可以满足推断模型所有权的目的。

5.2 评估近边界特性

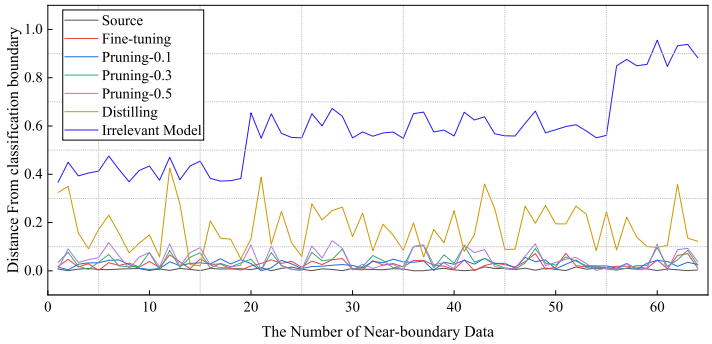
如图3所示, 本文提出的近边界数据在所有盗窃模型中都表现出了靠近分类边界的特性。例如，我们的方法在全部数据集中，近边界数据在无关模型上表现出和目标分类边界没有关系。然而，在基于源模型的各种盗窃模型上近边界数据表现良好，尤其是模型修改方法破坏性较强如模型蒸馏时依旧表现出靠近目标分类边界的特性。注意我们此处使用的近边界数据的大小为 64，换句话说，我们的方法在较小规模的近边界数据上依旧效果显著，可以想到的是当数据量增大时，我们的方法效果会更加明显。

5.3 Defending against Model Stealing

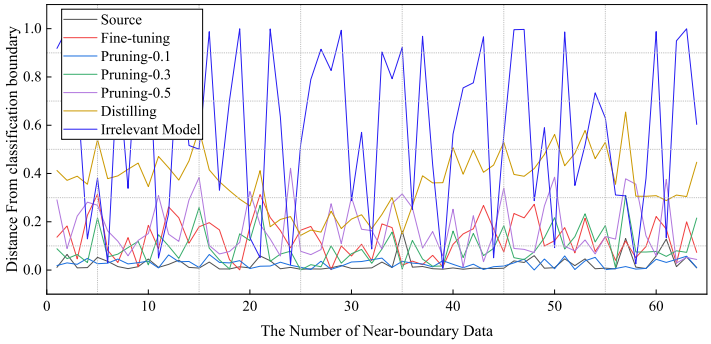
在本节中，我们讨论了我们方法在与其他近边界数据在盗窃模型上的性能对比，我们模拟了盗窃者可能会提供的近边界数据，该数据由几部分组成，包括 (1) 从原始数据中挑选出的近边界数据，(2) 由 FGSM 和 CW 生成的一些对抗性样本。我们针对不同的目标分类边界，按照 Section 4.3中进行假设检验并计算在不同数据集和不同盗窃模型上的 $\Delta\mu$ 和 p-value。如表5.3所示,the p-value 显著低于 0.05 在全部情况中。换句话说，我们的方法在不同的盗窃方法中推断模型所有权均有显著的效果。尽管在模型蒸馏上表现不如其他情况。实验结果表明使用私有的近边界数据对大多数模型盗窃技术都是可靠的，我们可以声称模型被盗



(a) CIFAR-10



(b) Heritage



(c) Intel_image

图 3: The performance of near-boundary data on different datasets and different model stealing methods.

窃的置信度至少为 95%，模型蒸馏是我们方法的最大挑战，也同样是其他研究面临的巨大挑战，在我们的实验中，可以观察到我们的模型始终可以将蒸馏模型标记为被盗的模型，这证明了我们方法的鲁棒性。

5.4 Scaling

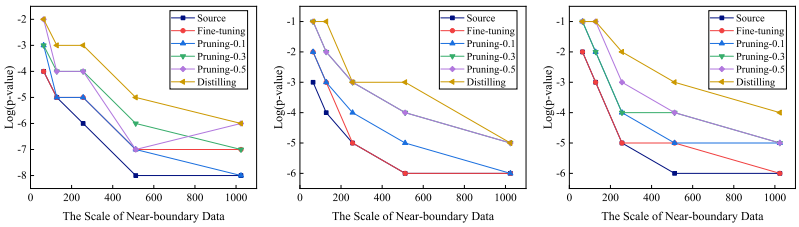
在本节中，我们测试了不同规模的近边界数据在推断模型所有权上的可伸缩性。回想我们的方法需要对数据进行采样从而进行假设检验，通常来说样本数量越大，对检验过程中因随机而产生的不利影响越少，在推断所有权上越有信心。如图4所示，p-value 随着近边界数据的规模增大而减少，也就是说近边界数据越多推断的信心也越多，但这并不说明我们的方法对小数量的近边界数据缺少鲁棒性，从表中我们可以观察到即使数据量为 64 的情况下，p-value 仍小于 0.05，这证明我们的方法对于小数量同样有显著的效果。

6 Realated Work

目前大多数的研究都是基于模型水印的方法，且都会通过绑定水印和模型性能的方法来抵御窃取威胁[]。然而大多数情况，在模型中增加噪声会导致模型的性能下降 [Maini, Yaghini, and Papernot 2021]。模型水印 [Co 57,DI>Jia 2020,Liu et al., 2018; Chen et al., 2019; Wang et al., 2019; Shafieinejad et al.Deepauth,

Dataset	Model Stealing	CB_1		CB_2		CB_3		CB_4		CB_5	
		$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value
CIFAR-10	Source	0.913	10^{-6}	0.954	10^{-6}	0.927	10^{-5}	0.967	10^{-5}	0.958	10^{-5}
	Fine-tuning	0.718	10^{-5}	**0.745**	10^{-6}	0.698	10^{-5}	0.692	10^{-4}	0.729	10^{-5}
	Pruning-0.1	**0.572**	10^{-5}	0.487	10^{-5}	0.458	10^{-5}	0.533	10^{-4}	0.512	10^{-4}
	Pruning-0.3	0.537	10^{-4}	0.497	10^{-4}	0.401	10^{-3}	0.428	10^{-4}	**0.587**	10^{-4}
	Pruning-0.5	0.545	10^{-4}	**0.614**	10^{-4}	0.506	10^{-3}	0.570	10^{-4}	0.484	10^{-3}
	Distilling	**0.372**	10^{-3}	0.297	10^{-3}	0.288	10^{-3}	0.308	10^{-3}	0.340	10^{-3}
Heritage	Source	0.876	10^{-5}	0.845	10^{-5}	0.859	10^{-4}	0.801	10^{-4}	0.837	10^{-5}
	Fine-tuning	0.815	10^{-5}	0.792	10^{-4}	0.824	10^{-4}	**0.833**	10^{-4}	0.784	10^{-4}
	Pruning-0.1	0.530	10^{-4}	**0.535**	10^{-3}	0.508	10^{-4}	0.486	10^{-3}	0.471	10^{-3}
	Pruning-0.3	**0.491**	10^{-3}	0.452	10^{-3}	0.469	10^{-4}	0.470	10^{-3}	0.427	10^{-4}
	Pruning-0.5	0.502	10^{-3}	**0.517**	10^{-3}	0.434	10^{-3}	0.451	10^{-3}	0.490	10^{-3}
	Distilling	0.329	10^{-3}	**0.365**	10^{-2}	0.238	10^{-3}	0.310	10^{-3}	0.274	10^{-3}
Intel_image	Source	0.859	10^{-5}	0.896	10^{-4}	0.872	10^{-4}	0.899	10^{-4}	0.914	10^{-4}
	Fine-tuning	0.717	10^{-5}	0.784	10^{-4}	0.752	10^{-4}	**0.791**	10^{-3}	0.709	10^{-4}
	Pruning-0.1	0.451	10^{-4}	0.522	10^{-4}	**0.539**	10^{-3}	0.472	10^{-3}	0.438	10^{-4}
	Pruning-0.3	0.407	10^{-4}	**0.415**	10^{-4}	0.346	10^{-3}	0.382	10^{-3}	0.395	10^{-3}
	Pruning-0.5	0.370	10^{-3}	0.395	10^{-3}	0.327	10^{-3}	0.360	10^{-3}	**0.458**	10^{-3}
	Distilling	0.336	10^{-2}	**0.395**	10^{-3}	0.360	10^{-2}	0.308	10^{-3}	0.287	10^{-2}

表 2: The performance of inferring model ownership for different target classification boundaries.



(a) CIFAR-10 (b) Heritage (c) Intel_image

图 4: The scalability of Inferring Model Ownership on Near-Boundary Data.

2019，shuiyin8]。水印的研究目的是验证模型的所有权而不是防御模型盗窃，嵌入和提取水印是模型水印的重要操作，例如将水印作为参数嵌入神经网络 [Co 57] 和设置触发器作为水印嵌入到模型中 [shuiyin2,Li et al. 2020b]. 此外，为了增强模型修改的鲁棒性，嵌入护照 [1,2,3] 的思路被提出，其操作是为模型增加特殊的护照层，在无法获取护照的情况下模型被盗的威胁被降低。水印被广泛使用作为验证所有权的一种方式，易对模型性能造成影响，一种从模型中提取知识作为指纹 [zhiwen0-3] 并验证所有权的想法被提出。研究者 [zhiwen0] 希望可以使用决策边界作为指纹来验证所有权，同样也有使用激活函数的性质 [zhiwen2] 抽象表示神经网络第二层的知识作为模型指纹。然而，无论是水印还是指纹都易受到歧义攻击 [Deepauth Fan, Ng, and Chan 2019; Li et al. 2019a]。歧义攻击旨在通过为 DNN 模型伪造其他水印或指纹来对所有权验证产生怀疑。直觉上，如果对手可以在水印模型上嵌入第二个水印或提取第二个指纹，那么该模型的 IP 所有权存在巨大的歧义。重要的是要注意，歧义性攻击是我们提出方法的关注。

针对模型盗窃开发的技术问题涉及到模型修改，包括模型微调，模型剪枝和模型压缩等。微调 [综述 1 2] 常见于迁移学习的过程中，涉及重新调整模型以更改模型的参数，同时保持性能。模型微调可以通过微调现有模型派生出许多模型。剪枝 [综述 4] 是部署 DNN 的常见方法，通过使用参数修剪来减少 DNN 的内存和计算开销，而盗窃者可能会使用修剪来删除水印或指纹。模型压缩 [综述 2] 中常见的是蒸馏，通过将大模型的知识蒸馏到小模型中可以显着降低内存需求和计算开销，现在的研究 [Hinton, Vinyals, and Dean 2014] 甚至不需要原始训练数据可以直接利用 API 蒸馏模型，因此常被用于派生模型。

7 Discussion and Conclusion

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

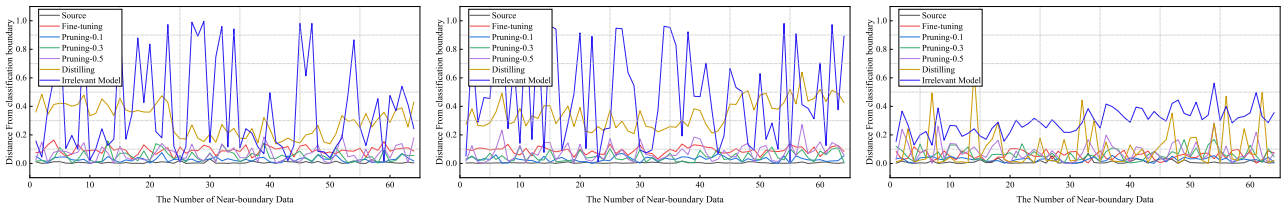
特性也会被被盗模型所继承。因此，我们构建了一种有趣的近边界数据用以推断所有权，并设计了使用 CW-L_2 迭代添加小噪声的方法生成对抗性样本，这是初始的近边界数据。我们训练了一种基于 DCGAN 的近边界生数据成器用以将近边界数据私有化和扩展，实验测试了生成器能够显著地学习近边界数据的特征并生成新的数据。最后，我们设计了新的损失函数微调源模型的分类边界，得到了最终版本的源模型和近边界数据。我们在 CIFAR-10，Heritage 和 Intel_image 数据集上进行评估，实验证明了我们的方法可以高置信度地推断模型所有权。

A 实验设置

本文实验利用 CIFAR-10，Heritage 和 Intel_iamge 三种开源数据集用于 ResNet18，训练过程中使用 Adam 优化器并将 Learning rate，Epoch and 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 $\alpha = 0.7$, 训练 epoch 是 20。本文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 Learning rate，迭代次数和二分搜索次数分别被设置为 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，训练过程中使用 Adam 优化器且将学习率，训练轮次和 Batch size 分别设置为 0.0002，8000 和 64。注意本文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分类边界的损失函数来微调源模型，具体设置为每 10 个轮次交替一次且交替次数最多为 10 次。

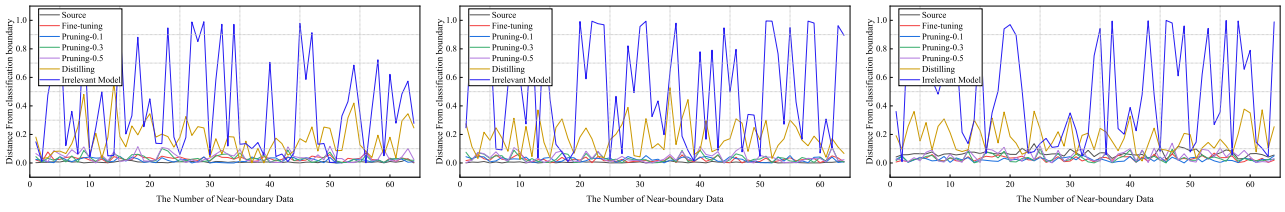
B 评估近边界特性的扩展

本小节将对 Section 5.2 中实验进行扩展。In Section 5.2，本文只展示了数据规模为 64 时近边界数据在源模型、可疑模型和无关模型上的近边界特性，且只针对其中一条分类边界。所以，本小节将从不同分类边界的角度对该实验进行扩展。如图5,6 and 7所示, 本文所提出的近边界数据在不同的目标分类边界时都展示出显著的近边界特性，且源模型和可疑模型 (被盗模型) 之间的差异很小，而以上两者与无关模型之间的差异却较大。总之，近边界数据的近边界特性确实可以被用于推断模型所有权。



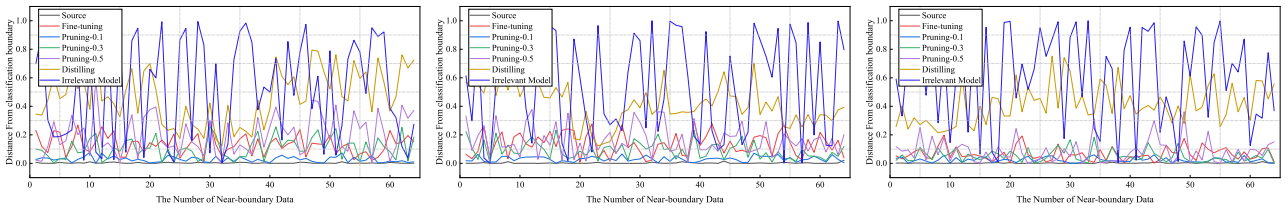
(a) CB_1 (b) CB_2 (c) CB_3

图 5: The performance of near-boundary data on CIFAR-10 and different model stealing methods for the different classification boundaries.



(a) CB_1 (b) CB_2 (c) CB_3

图 6: The performance of near-boundary data on Heritage and different model stealing methods for the different classification boundaries.



(a) CB_1 (b) CB_2 (c) CB_3

图 7: The performance of near-boundary data on Intel_image and different model stealing methods for the different classification boundaries.

Dataset	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**	**0.365**	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_images	0.522	0.447	**0.353**	**0.088**
	0.475	0.506	0.387	0.122
	0.468	**0.402**	0.428	0.127

表 3: 不同对抗性样例生成算法生成的数据与目标分类边界的平均距离（粗体为平均距离最小值）.

C 初始近边界数据的算法选择

本小节将对 Section 4.1 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 则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 Appendix A 所示。如3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。

D 微调目标分类边界的精度影响

本小节对 Section 4.2 中测试微调对源模型的影响进行了广泛的实验。我们针对不同数据集训练得到的源模型，使用不同规模的近边界数据以及不同的目标分类边界对源模型进行微调。如4，5 and 6所示，几乎全部情况下的源模型微调前后精度差均不超过 5pp。实验结果证明，我们的近边界数据在应用于模型所有权保护问题中，具有较好的保真度。因此，使用近边界数据推断模型所有权不用担心源模型受到较大的影响。

E 可伸缩性的扩展

Section 5.4 中我们测试了近边界数据推断模型所有权在可伸缩性上的表现。本小节在其基础上扩展为针对不同分类边界的可伸缩性测试。如图**??**所示，对于不同数据集训练得到的源模型和可疑模型，我们的根据不同目标分类边界生成得到的近边界数据在推断模型所欲权上均表现出显著的效果，与 Section 5 中得到的结果相同，这极大地支持了我们提出的方法。

Dataset	CB	Data Size	Acc.
CIFAR-10	CB_1	64	0.873
		128	0.862
		256	0.862
		512	0.854
	CB_2	64	0.871
		128	0.870
		256	0.860
		512	0.844
	CB_3	64	0.871
		128	0.868
		256	0.858
		512	0.856

表 4: 使用不同规模的近边界数据微调三条目标分类边界后源模型的精度（微调前为 0.886）.

Dataset	CB	Data Size	Acc.
Intel_image	CB_1	64	0.755
		128	0.769
		256	0.756
		512	0.779
	CB_2	64	0.770
		128	0.741
		256	0.768
		512	0.777
	CB_3	64	0.781
		128	0.753
		256	0.764
		512	0.752

表 5: 使用不同规模的近边界数据微调三条目标分类边界后源模型的精度（微调前为 0.879）.

Dataset	CB	Data Size	Acc.
Heritage	CB_1	64	0.856
		128	0.825
		256	0.830
		512	0.797
	CB_2	64	0.823
		128	0.839
		256	0.841
		512	0.779
	CB_3	64	0.848
		128	0.826
		256	0.779
		512	0.791

表 6: 使用不同规模的近边界数据微调三条目标分类边界后源模型的精度（微调前为 0.794）.

F Rebuttal

感谢 reviewer 们用心的 review，我将对提出的 review 给出 reponse。[Reviewer #1.]Q1.) 本文的假设其实很明确，因为白盒或者黑盒的定义是指在验证 (推断) 所有权的时候是否需要访问模型内部信息。本文提出的方法在模型训练时部署并生成近边界数据。当受害者提出对可疑模型的质疑时，只需输入近边界数据观察输出结果，无需访问模型内部信息。因此，本文提出的是方法是基于黑盒环境的。Q2.）Cao et al.[2] 和 Heo et al.[3] 提出使用对抗性样本表示决策边界的方法反映模型知识。需要注意的是：(1) 对抗性样本虽然距离决策边界较近，但本文提出的近边界数据需要的是“最近”的数据，单纯的对抗性样本不能满足推断模型所有权的要求。(2) 对抗性样本的使用通常是与正常样本成对出现的，两者共同工作产生作用。然而，本文的近边界数据只是一种最靠近决策边界的数据，两者使用方法也不同。Q3.）本文在 Section 3 与数据集推断方法进行比较，表明了数据集推断方法的局限性在于训练数据集必须是私有数据集，本文解决了这个局限。此外，本文提出的窃取方法已经超过了大多数已有工作。与数据集推断相比，本文将细化的盗窃方案整合，以适应篇幅限制。

[Reviewer #2.]Q1.)Reviewer 并没有完全理解我们的工作。我们提出的近边界数据不是证明可疑模型是被盗模型，而是我们提出的近边界数据比可疑者提供的近边界数据更靠近决策边界，通过推断我们可以声称暂时拥有所有权。我们不在乎窃取者的否认，只需比对方的近边界数据更靠近决策边界即可，我们利用了一种极限比较推断的思想。这是我们工作最大的创新点，我们是第一个工作提出将推断所有权替代验证所有权。reviewer 提到的水印工作即是验证所有权工作的例子，面临着歧义攻击，水印去除等威胁。Q2.）我们详细解释了本文的威胁模型参考 Reviewer #1-Q1。

[Reviewer #6.]Q1.) 我们的训练数据是公开的，但我们的近边界数据是私有的。这是我们比现有工作的先进性。Q2 and Q3.)reviewer 对我们提出方法有误解。我们的方法并没有使用对抗性样本作为指纹，也

不是一个指纹工作，而是利用对抗性样本生成近边界数据，再利用 DCGAN 将其私有化。这是一种使用数据作为推断所有权依据的想法。指纹是指挖掘模型内部固有信息，本文并没有这样做。Q4 and Q5.) 歧义攻击是指提出另一种水印或指纹重复声明模型所有权，而本文是使用近边界数据比较距离决策边界的距离，距离更近的被推断拥有所有权。参考 Section 4.3，只要提前与第三方机构备案决策边界，就可以避免歧义攻击。此外，关于 Weakness 2，攻击者的确可以通过访问模型，按照本文设计方法重新生成近边界数据，但存在两个问题：1) 必须通过 DCGAN，否则视觉上有差异由于数据集公开，2) 由于不知道约定的分类边界，需要大量的测试，这会导致极大的成本代价。我们知道，不存在无法攻击的防御方法，但可以从攻击成本上约束。

[Reviewer #7.]Q1.) 我们给出了详细解释，请参考 Reviewer #6-Q2。Q2.) 我们使用 DCGAN 的目的是将对抗性攻击生成的样本私有化。我们希望用于比较的数据能够和训练数据被区分，从而提升攻击者的攻击难度。我们将在修改的论文中增加视觉示例。Q3.) 本文在 Section 3 中提到了 [2][3] 的局限性，其中由于 [2] 要求训练数据的私有化，[3] 则是验证所有权的传统思路，与本文提出的方法难以在一个尺度内进行比较。然而，本文强制忽略上述问题，增加了比较实验，结果表明本文的方法在效果上和 [2][3] 等同。如果允许我们可以补充这部分实验。Q4.) 蒸馏是最难的挑战对于所有权工作，例如在 [2][3] 中在蒸馏上的效果也弱于其他盗窃技术。本文的方法同样对于蒸馏的效果弱于其他，但本文创新性地提出推断所有权代替验证所有权，因此只要受害者的近边界数据更靠近决策边界就可以声称拥有所有权。Q5.) 本文提出的方法尽管也会对模型有影响，但由于只微调其中很小一部分甚至一条，造成的影响远小于现有方法。如果允许我们可以补充实验结果以证明这一结果。Q6.)CB 是指分类边界，n 指已约定分类边界标号，我们将会解释完整。

Thanks to the reviewers for your attentive reviews, and I will give my responses. [Reviewer #1.] Q1.)The assumptions in this paper are actually clear, because the definition of the white or black box refers to whether access to the internal information of the model is required when verifying (inferring) ownership. The method proposed in this paper deploys and generates near-boundary data during model training. When a victim questions a suspicious model, they only need to input the near-boundary data to observe the output results without accessing the internal information of the model. Therefore, our method is based on the black-box environment. Q2.) Cao et al.[2] and Heo et al.[3] propose to reflect model knowledge using adversarial samples to represent decision boundaries. It is important to note that (1) although the adversarial samples are close to the decision boundary, our near-boundary data requires the "nearest" data, and the adversarial samples alone cannot meet the requirement of inferring the model ownership，(2) the use of adversarial samples usually occurs in pairs with normal samples, and the effect of both working together. However, our near-boundary data is just a kind of data closest to the decision boundary, so the two methods are totally different. Q3.) This paper compares with the Dataset Inference method in Section 3 and shows that the DI limitation is the training dataset must be private, and we solve this limitation. In addition, the stealing techniques proposed in this paper have surpassed most of the existing work. In contrast to DI, this paper integrates the refined stealing scheme to fit the page limitation.

[Reviewer #2.]Q1.)The reviewer has not fully understood our work. The near-boundary data we proposed is not to prove that the suspect model is the stolen model, but our near-boundary data is closer to the decision boundary than the data provided by the suspect, and by inference we can claim temporary ownership. We do not care about the denial of the stealer, we just need to be closer to the decision boundary than the other attackers, and we exploit an idea of limit comparison inference. This is the most significant innovation point of our work, and we are the first work to propose inferring ownership instead of verifying ownership. The reviewer mentions watermarking work that is an example of verifying ownership work, facing threats such as ambiguity attacks, watermark removal, etc. (Q2.) We explain our threat model in detail refer to reviewer #1-Q1.

[Reviewer #6.]Q1.) Our training data is publicly available, but our near-bound data is private. This is our advancement over existing works. Q2 and Q3.) The reviewer misunderstood our proposed method. Our method does not use adversarial samples as fingerprints and our work is not a fingerprinting work, but uses adversarial samples to generate near-boundary data and then uses DCGAN to make it private. This is an idea of using data as a basis for inferring ownership. Fingerprinting refers to mining the information inherent within the model. Q4 and Q5.) Ambiguity attack refers to proposing another watermark or fingerprint to repeatedly declare model ownership, while this paper uses near-boundary data to compare the distance from the decision boundary, and the closer one is inferred to have ownership. Referring to Section 4.3, the ambiguity attack can be avoided by filing the decision boundary with a third-party organization in advance. Moreover, regarding Weakness 2, an attacker can indeed regenerate the near-boundary data by accessing the model and following the design method in this paper, but there are two problems: 1) it must pass DCGAN, otherwise there is a visual discrepancy due to the public dataset, and 2) it requires a lot of testing because the agreed classification boundary is not known, which can lead to a tremendous cost. We know that no defense method cannot be attacked, but it can be constrained in terms of the attack cost.

[Reviewer #7.]Q1.) We give a detailed explanation, please refer to Reviewer #6-Q2. Q2.) We use DCGAN to privatize the samples generated by the adversarial attack. We expect the data used for comparison to be distinguished from the training data, thus improving the attacker's difficulty of attack. We will add visual examples in the revised paper. Q3.) This paper mentions the limitations of [2,3] in Section 3, where it is difficult to compare with our method within a scale because [2] requires the privatization of the training data and [3] is a traditional idea of verifying the ownership. However, we force to add comparative experiments, and the results show that our method is equivalent to [2,3]. We can add this experiment if allowed. Q4.) Distillation is the most difficult challenge for ownership work, for example in [2,3], where the effect on distillation is also weaker than others. Our method is similarly weaker for distillation than others, but we innovatively proposes to infer ownership, so ownership can be claimed as long as the victim's near-boundary data is closer to the decision boundary. Q5.) Although our method also impacts the model, it causes much less impact than existing methods because only a tiny fraction is fine-tuned. If allowed we can add experimental results. Q6.) CB refers to the classification boundary and n refers to the agreed classification boundary notation, which we will explain in full.