计算机科学
COMPUTER SCIENCE

针对机器学习的成员推断攻击综述

彭钺峰，赵波，刘会，安杨

引用本文

彭钺峰，赵波，刘会，安杨.针对机器学习的成员推断攻击综述[J]．计算机科学, 2023, 50(3): 351-359.

PENG Yuefeng, ZHAO Bo, LIU Hui, AN Yang. Survey on Membership Inference Attacks Against Machine Learning [J]. Computer Science, 2023, 50(3): 351-359.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100016

针对机器学习的成员推断攻击综述

彭钺峰[1]　赵　波[1]　刘　会[1]　安　杨[2]
1 武汉大学国家网络安全学院　武汉 430000
2 武汉大学计算机学院　武汉 430000
(yuefengpeng@whu.edu.cn)

摘　要　近年来,机器学习不仅在计算机视觉、自然语言处理等领域取得了显著成效,也被广泛应用于人脸图像、金融数据、医疗信息等敏感数据处理领域。最近,研究人员发现机器学习模型会记忆它们训练集中的数据,导致攻击者可以对模型实施成员推断攻击,即攻击者可以推断给定数据是否存在于某个特定机器学习模型的训练集。成员推断攻击的成功,可能导致严重的个人隐私泄露。例如,如果能确定某个人的医疗记录属于某医院的数据集,则表明这个人曾经是那家医院的病人。首先介绍了成员推断攻击的基本原理;然后系统地对近年来代表性攻击和防御的研究进行了总结和归类,特别针对不同条件设置下如何进行攻击和防御进行了详细的阐述;最后回顾成员推断攻击的发展历程,探究机器学习隐私保护面临的主要挑战和未来潜在的发展方向。

关键词:机器学习;成员推断;隐私泄露;隐私保护
中图法分类号　TP181

Survey on Membership Inference Attacks Against Machine Learning

PENG Yuefeng[1],ZHAO Bo[1],LIU Hui[1] and AN Yang[2]
1 School of Cyber Science and Engineering,Wuhan University,Wuhan 430000,China
2 School of Computer Science,Wuhan University,Wuhan 430000,China

Abstract　In recent years,machine learning has not only achieved remarkable results in conventional fields such as computer vision and natural language processing,but also been widely applied to process sensitive data such as face images,financial data and medical information. Recently,researchers find that machine learning models will remember the data in their training sets,making them vulnerable to membership inference attacks,that is,the attacker can infer whether the given data exists in the training set of a specific machine learning model. The success of membership inference attacks may lead to serious individual privacy leakage. For example,the existence of a patient's medical record in a hospital's analytical training set reveals that the patient was once a patient there. The paper first introduces the basic principle of membership inference attacks,and then systematically summarizes and classifies the representative research achievements on membership inference attacks and defenses in recent years. In particular,how to attack and defend under different conditions is described in detail. Finally,by reviewing the development of membership inference attacks,this paper explores the main challenges and potential development directions of machine learning privacy protection in the future.

Keywords　Machine learning,Membership inference,Privacy leakage,Privacy protection

1　引言

近年来,机器学习在计算机视觉[1-2]、自然语言处理[3-4]和语音识别[5-6]等多个领域都取得了显著的发展。机器学习因其强大的性能而被广泛应用于现实数据的处理任务中,例如人脸识别[7-8]、医疗分析[9-10]等。最近的研究表明,机器学习模型能够记忆训练数据的敏感信息[11-13],使它们容易受到不同类型的隐私攻击,包括模型窃取攻击[14-16]、模型逆向攻击[17-19]以及最值得注意的成员推断攻击[20-22].

在成员推断攻击中,给定一个数据样本(如一张人脸图像或一个医疗记录),攻击者可以推断这个样本是否参与目标模型的训练。根据 Yeom 等[22]的研究,成员推断攻击背后的原理是:对于机器学习模型来说,当输入的数据是它训练集中的数据样本(即成员)时,模型的行为通常与输入为它们第一次"看到"的数据(即非成员)时不同。成员推断攻击的一般流程如图 1 所示,攻击者首先将样本点作为输入访问目标模型,然后根据目标模型的输出判断该样本是否是目标模型训练集的成员。一次成功的成员推断攻击,可能会给机器学习模型的

到稿日期:2022-01-04　返修日期:2022-03-27
基金项目:国家自然科学基金(U1936122)
This work was supported by the National Natural Science Foundation of China(U1936122).
通信作者:赵波(zhaobo@whu.edu.cn)

拥有者带来严重的隐私风险。例如，攻击者能够确定某人的医疗记录参与了医院机器学习模型的训练，则可以推断出这个人曾经是该医院的病人。

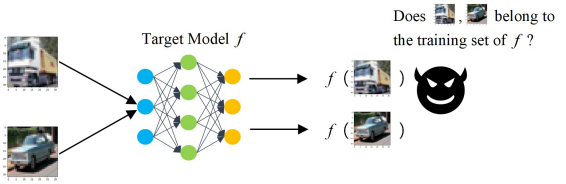

图 1 成员推断攻击

Fig. 1 Membership inference attacks

机器学习在现实中被广泛应用，主流云提供商如谷歌(Google)、亚马逊(Amazon)等已经部署了机器学习即服务(MLaaS)，允许客户上传自己的数据并为他们提供模型训练服务。客户还可以将经过训练的模型挂到平台上进行租赁，供给外部用户进行在线使用并收取单次访问费用，具体流程如图 2 所示。MLaaS 平台使得攻击者伪装成良性用户访问机器学习模型成为可能。

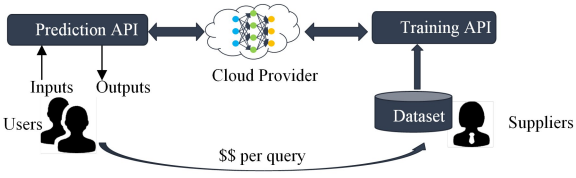

图 2 机器学习即服务

Fig. 2 Machine learning as a service

在机器学习服务迅速商业化的背景下，成员推断攻击造成的隐私风险会极大地阻碍机器学习模型的商业化部署，甚至可能导致提供机器学习服务的企业违反隐私条例。因此，成员推断攻击的研究对保护机器学习的隐私具有重大意义。

Shorki 等[20]于 2017 年首次提出了针对机器学习的成员推断攻击，并在谷歌的 MLaaS 平台上进行实验，证明了在此类服务下存在以外部用户的身份进行攻击的可能。此后，成员推断攻击在更多的场景下被证明可行。与此同时，很多防御性工作也被提出用于缓解这种攻击。尽管成员推断的攻击和防御是一个新兴且正在快速发展的研究领域，但是目前这一领域仍然缺少系统性的总结和综述。

本文首先从成员推断攻击的原理出发，分别从攻击和防御两个角度重点调研了这一领域的关键技术和代表性工作，并对这些工作进行了系统性的总结和归类。通过回顾成员推断攻击的发展以及讨论现有工作的局限性，指出未来成员推断攻击的潜在发展方向。本文的主要贡献如下：

(1)形式化定义了成员推断攻击，并从攻击者拥有的对抗性知识等方面入手，详细阐述了成员推断攻击在白盒、灰盒和黑盒场景下的威胁模型，提供了该领域的整体概述；

(2)系统地调研了成员推断攻击和防御领域的代表性工作，并对这些方法进行了细致的阐述和全面的梳理，完善了成员推断领域的方法论总结；

(3)从攻击和防御两个角度，总结并讨论了成员推断现有工作的不足和限制，并指出未来可能的发展方向。

2 背景知识

2.1 机器学习

成员推断攻击的目标为有监督分类机器学习，本文主要讨论这类机器学习任务。这些任务要求机器学习模型学习如何为给定的数据点 $\boldsymbol{x}$ 划分正确的类别，即给数据点 $\boldsymbol{x}$ 打上正确的类别标签 y。通常 $\boldsymbol{x}$ 是多维特征向量，表示一个图像或一个句子，y 是对应于 $\boldsymbol{x}$ 的标签。机器学习模型本质上可以被看成一个映射函数 f，该函数以 $\boldsymbol{x}$ 作为输入，并输出一个向量 $f(\boldsymbol{x};\theta)$，其中 θ 是 f 的参数。

模型的训练原理如下：给定一个数据集 $D=\{\boldsymbol{x}^n,y^n\}_{n=1}^N$，对于数据集中的每个样本 $\boldsymbol{x}$，模型都会给出对应的输出 $f(\boldsymbol{x};\theta)$。模型输出 $f(\boldsymbol{x};\theta)$和真实标签 y 的距离可以用一个函数 $\mathscr{L}$ 表示，这个函数被称为损失函数。模型在数据集上的总体损失被称为经验风险，具体定义如式(1)所示。模型训练的目标就是找到一组最优参数 θ^*，使得模型在 D 上的经验风险最小，优化目标如式(2)所示。

$$\mathscr{R}_D(\theta)=\frac{1}{N}\sum_{n=1}^{N}\mathscr{L}(f(\boldsymbol{x}^n;\theta),y^n) \tag{1}$$

$$\theta^*=\arg\min_{\theta}\mathscr{R}_D \tag{2}$$

训练完成的模型可以用来对新的样本做决策。以有 K 个类的分类问题为例，对于一个新的样本 $\boldsymbol{x}$，模型输出一个包含 K 个置信度的预测向量 $f(\boldsymbol{x};\theta)$，所有置信度之和为 1。置信度最高的类则被预计为 $\boldsymbol{x}$ 的标签，即 $y=\arg\max f_i(\boldsymbol{x};\theta)$。模型的训练和预测过程如图 3 所示。

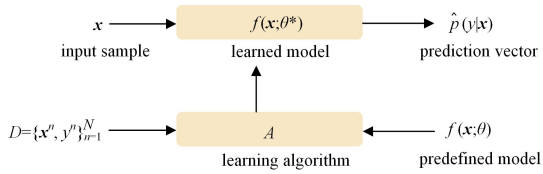

图 3 典型机器学习过程

Fig. 3 Typical machine learning process

机器学习模型通过优化算法从大量历史数据(数据集)中学习规律，从而对新的样本做决策。但是另一方面，由于模型的参数是基于训练集中的样本进行优化的，这可能导致模型产生过拟合现象，即模型在训练集上的性能优于模型在其他数据上的性能[23-25]。有研究表明，过拟合是导致成员推断攻击的直接原因之一[22]，我们将在第 6 节中对此进行具体分析，并从机器学习的角度讨论其他影响成员推断攻击的因素。

2.2 成员推断攻击

在针对机器学习的成员推断攻击中，攻击者的主要目的是判断一个给定的样本$(\boldsymbol{x},y)$是否在目标模型 θ 的隐秘训练数据集 D_{tr}中。根据这个目的，成员推断攻击可以被看作一个二分类问题：把样本分类为成员或者是非成员。我们首先给出通用的成员推断攻击定义：

$$\mathscr{A}:(\boldsymbol{x},y),\theta,\mathscr{K}\rightarrow\{0,1\} \tag{3}$$

其中，$\mathscr{K}$ 代表攻击者对目标模型的了解情况，1 代表$(\boldsymbol{x},y)$是 D_{tr}的成员，0 代表非成员。

在成员推断攻击中，攻击者通常训练一个二分类器 A 来解决上述二分类问题，这个二分类器被称为攻击模型。训练

攻击模型时的增益函数如式(4)所示。

$$G_\theta(A)=\mathbb{E}_{(\boldsymbol{x},y)\sim \mathbf{Pr}_{D_{\mathrm{tr}}}}[\log A(\theta(\boldsymbol{x}),y)]+\mathbb{E}_{(\boldsymbol{x},y)\sim \mathbf{Pr}_{\backslash D_{\mathrm{tr}}}}[1-\log A(\theta(\boldsymbol{x}),y)] \tag{4}$$

其中，$\mathbf{Pr}_{D_{\mathrm{tr}}}$ 和 $\mathbf{Pr}_{\backslash D_{\mathrm{tr}}}$ 分别是 D_{tr} 中的样本和 D_{tr} 外的样本的分布。攻击模型训练成功后，攻击者即可以用它来对新的样本预测成员。对于一个样本$(\boldsymbol{x},y)$，攻击者首先将其输入目标模型 θ 中获得输出 $\theta(\boldsymbol{x})$，随后攻击二分类器以$(\theta(\boldsymbol{x}),y)$为输入判断样本是否是成员。在具体的攻击方法中，该攻击二分类器可以是一个机器学习模型(如神经网络)[20-21]。对于每个样本，该机器学习模型计算出样本是成员的概率为 $A(\theta(\boldsymbol{x}),y)$。如果 $A(\theta(\boldsymbol{x}),y)>0.5$，则攻击者认为 $\boldsymbol{x}$ 是 D_{tr} 的成员。攻击二分类器也可以是一个精心调整的 Metric 阈值，例如样本的 loss 值[22]、置信度[21]或样本到模型决策边界的距离[26]等。攻击首先计算样本对应的 Metric 值，再根据该值是否超过(低于)阈值来判断该样本是成员还是非成员。第 4 节将详细描述如何使用不同的攻击二分类器进行成员推断。

同时，根据对目标模型了解程度的不同，攻击者可以采取不同的方案训练这些攻击模型。例如，在拥有足够多目标模型知识的情况下，攻击者可以训练影子模型来模拟目标模型，并使用影子模型来辅助二分类攻击模型的训练[20-21]或是攻击阈值的调整[22,26]。第 3 节具体介绍攻击者在攻击中可能使用的目标模型知识。

2.3 数据集

成员推断攻击和防御的研究通常需要在相同的数据集上评估方案的性能，以提供比较和参考。常用的数据集包括机器学习领域的通用数据集，如 MNIST 和 CIFAR-10[27]；也包括专门为成员推断设计的数据集，如 Purchase，Texas 和 Location。MNIST 数据集来自美国国家标准与技术研究所，包含 60000 个训练样本和 10000 个测试样本，每个样本都是一张手写数字的图片(0-9)。CIFAR-10 是 Alex Krizhevsky 等搭建的小型图像数据集，包括 50000 张训练样本和 10000 个测试样本。这些样本共分 10 类：飞机、汽车、鸟、猫、鹿、狗、蛙、马、船和卡车。MNIST 和 CIFAR 数据集内容简单，尺寸小，通用性好，因此经常被用于机器学习多个领域的实验。

Purchase，Texas 和 Location 是成员推断攻击的提出者 Shokri 等为成员推断专门构建的数据集[20]。Purchase 基于 Kaggle 上收集的用户购买数据，原始数据集包含大量用户一年内的购买记录，但是不具有任何分类。Shokri 等对原始数据集应用聚类，将这些数据分成 100 个类，分别对应不同的购物模式。Location 数据集基于大量曼谷用户在一年时间里的旅行记录生成，被划分成 30 类。每条数据包含 446 个二进制特征，分别代表该用户是否曾去过某个地点。类似地，Texas 数据集由 Shokri 创建，收集了多家医疗机构患者的住院信息，每条数据包含 6170 个特征和 100 个分类，每个分类代表 1 种特定的医疗流程。

3 成员推断攻击分类

给定一个目标模型 θ 以及目标模型的训练集 D_{tr}，攻击者可以根据对 θ 和 D_{tr} 了解程度，设计不同的成员推断攻击方案。在成员推断攻击中，攻击者掌握的有关目标模型和目标数据集的知识被称为对抗知识。根据攻击者拥有的对抗知识的不同，成员推断攻击可以分为白盒攻击、黑盒攻击和灰盒攻击 3 类。下面首先介绍成员推断攻击中的对抗知识，然后给出按照对抗知识分类成员推断攻击的方法。

3.1 对抗知识

(1)数据知识

数据知识指攻击者对目标模型的训练数据分布的了解情况。掌握完全的数据知识会使成员推断攻击失去意义，因为攻击者已经知道了哪些是 D_{tr} 的成员。因此，成员推断攻击通常假设攻击者没有或者只具有部分 D_{tr} 的数据知识。攻击者可以获得和 D_{tr} 同分布或者分布相近的数据点，利用这些数据点，攻击者可以更好地训练攻击模型。

(2)训练知识

训练知识指攻击者对目标模型训练过程的了解程度，其主要包括训练目标模型使用的优化算法和对应的超参数配置等。以神经网络为例，训练知识包括训练神经网络使用的优化算法(如 SGD，Adam 等)和训练时使用的超参数配置(如学习率配置、训练 epoch 和 batch 大小等)。

这些知识揭示了模型具体是如何被训练的，知道这些知识意味着攻击者可以更好地分析目标模型的行为，识别目标模型在训练集和测试集上不同的行为，从而实施成员推断攻击。例如，攻击者可以在拥有训练知识的情况下，使用和目标模型相同的优化算法和参数配置训练影子模型来模拟目标模型，并通过观察影子模型在训练集和测试集上的不同行为来为攻击目标模型提供参考。

(3)模型知识

模型知识指目标模型自身的信息，包括目标模型的结构和模型内部的参数。目标模型的结构知识包括目标模型的类型(如神经网络)，以及目标模型内部的架构(如该神经网络的层数、内部激活函数的类型)。模型内部参数的知识指模型训练后内部的参数值(如经过训练后，神经网络中每个神经元的权重和偏置的值)。掌握模型知识的程度，往往决定了攻击者能实施的攻击类型。掌握模型参数的知识，意味着攻击者拥有对模型完全的访问权限，可以监测模型的行为，从而实施更强力的攻击。

(4)输出知识

输出知识指模型输出的预测向量 $\theta(\boldsymbol{x})$ 对于攻击者的可用程度，包括全部向量、部分向量和仅-标签 3 种程度。全部输出知识意味着，对于任何一个样本点 $\boldsymbol{x}$，攻击者拿到模型对于它的完整输出向量 $\theta(\boldsymbol{x})$；部分输出知识意味着，攻击者只能获取输出向量的部分值，例如每个输出向量中最大的 3 个值；仅-标签知识表示攻击者只能知道输出向量中拥有最大置信度的类别是哪一个，而不能获取任何置信度的值。3 种程度对应了 3 种不同难度的成员推断攻击，仅-标签的攻击难度最高，因为攻击者只能获取模型输出的标签，并通过输出标签是否正确来尝试推断成员。

3.2 成员推断攻击类别

如表 1 所列，我们根据攻击拥有对抗知识的不同程度，把成员推断攻击分成黑盒攻击、白盒攻击和灰盒攻击 3 类。数据知识和输出知识对于大部分成员推断攻击都是可用的，而 3 类攻击的差别主要体现在攻击者对模型知识和训练知识的了解上。

表 1 成员推断攻击的分类

Table 1 Taxonomy of membership inference attacks

攻击场景	数据知识	训练知识	模型知识	输出知识
黑盒攻击	无	无	无	全部/部分
灰盒攻击	部分	全部	部分	全部/部分
白盒攻击	部分	全部	全部	全部

白盒攻击假设攻击者具有完全的模型知识和训练知识，包括模型的类型、架构和参数。攻击者可以把数据输入模型，并直接观察目标模型各方面的行为，包括获取模型最后输出的置信度、模型的梯度值、loss 值，以及模型计算时的中间结果(如神经网络隐藏层的输出)。

灰盒攻击假设攻击者具有完全的训练知识和部分的模型知识。攻击者知道模型是如何训练的，且知道模型的类型和结构，但是不知道目标模型内部的参数。在这种设置下，攻击者把需要判断的样本输入模型后，无法直接获取目标模型内部的中间输出，只能获取最后的输出结果。但是攻击者可以用已知的知识训练出一个和目标模型相近的影子模型，以提供更多参考。

黑盒攻击假设攻击者不知道目标模型的训练知识和模型知识。在这种场景下，目标模型对于攻击者来说只是一个黑盒的 API，攻击者仅能把样本输入模型，并获取模型的输出，无法观察任何内部中间结果。黑盒的攻击的难度最高，因为攻击者对于目标模型来说只是一个普通的用户，不能像白盒攻击一样拿到模型内部的参数，也无法像灰盒攻击一样训练一个和目标模型相近的模型来获得参考。

4 关键技术研究进展

成员推断攻击的发现源自研究者对机器学习应用中隐私问题的不断探究。Shokri 等[20]发现 Google 和 Amazon 等部署的 MLaaS 服务存在隐私漏洞，攻击者可以通过服务提供的 API 访问训练好的机器学习模型，并推测某条数据是否在这个模型的训练集中，从而首次提出了针对机器学习的成员推断攻击。这个发现促进了研究人员对机器学习中成员信息隐私问题的思考和对各种场景和条件下攻击的探索。本节按提出的时间顺序介绍近年来成员推断攻击的代表性成果，并按照第 3 节所介绍的分类方法对这些攻击方案进行分类，给出攻击需要的详细设置，具体如表 2 所列。

表 2 代表性成员推断攻击

Table 2 Representative membership inference attacks

攻击方法		攻击类型	数据知识	模型知识	训练知识	输出知识
Shadow Training 攻击		灰盒	数据集的分布	模型类型	优化算法+超参数配置	全部输出向量
Metric 攻击	基于 correctness	黑盒	无	无	无	仅-标签
	基于 loss	灰盒	数据集分布	模型类型	优化算法+超参数配置	输出向量的 loss
NSH 攻击		白盒	数据集分布	模型类型+内部参数	优化算法+超参数配置	全部输出向量
ML-Leaks 攻击	Adversary I	灰盒	数据集分布	模型类型	优化算法+超参数配置	Top-3 输出向量
	Adversary II	黑盒	无	无	无	Top-3 输出向量
	Adversary III	黑盒	无	无	无	Top-1 输出向量
BlindMI 攻击		黑盒	无	无	无	Top-2 输出向量
Label-only 攻击		灰盒	数据集分布	模型类型	优化算法+超参数配置	仅-标签

4.1 Shadow Training 攻击

Shokri 等[20]发现了 MLaaS 服务可能存在的隐私风险，并针对其场景设计了第一个面向机器学习模型的成员推断攻击。如第 2 节提到的，成员推断攻击可以被看作一个二分类问题，Shorki 等提出训练一个神经网络模型作为攻击模型来进行二分类。给定任意样本数据，攻击模型返回该样本是成员的概率，如果该概率高于 0.5，则认为样本属于成员。训练这样的一个攻击模型需要收集一定数量有成员标签的训练数据，也就是说，攻击者需要知道这些训练数据是成员还是非成员。攻击者无法直接通过目标模型构建一个有标签的训练集。为了解决这个问题，Shokri 等提出了 Shadow Training 技术，使用与目标模型一样的模型结构和训练方法，训练多个影子模型来模仿目标模型的行为，具体流程如图 4 所示。

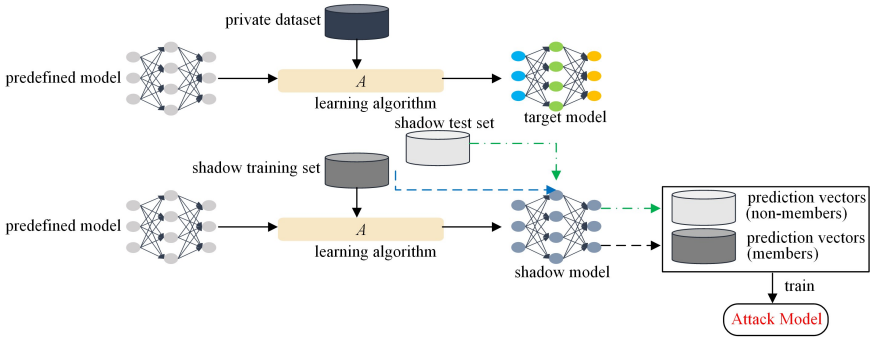

图 4 影子训练技术

Fig. 4 Overview of shadow training

影子模型由攻击者训练，因此攻击者可以根据影子模型的训练数据分配成员标签，从而构建一个有标签的训练集以训练攻击模型。Shadow training 技术给攻击者提供了训练攻击模型时的参考，是很多攻击方法的基石。

Shorki 等提出的攻击是基于影子模型的迁移攻击，这种方法的劣势是：当影子模型和目标模型差别很大时，攻击效果不佳。而训练一个高质量的影子模型，需要知道目标模型的全部训练知识和模型知识，无法在黑盒场景下实施。

4.2 基于 Metric 的攻击

Yeom 等[22]进一步分析了成员推断攻击的成因，认为过拟合是导致模型成员信息泄露的根本原因之一：模型在训练集（成员）上表现出的性能优于在测试集（非成员）上的性能，并且在训练集上的输出和在测试集上的输出具有不同的数值特征。他们基于这种观点，提出了 2 种基于 Metric 阈值的攻击，分别基于模型的 correctness 和 loss 值。

基于 correctness 的攻击利用目标模型在训练集和测试集上的准确性差距。具体来说，对于一个样本，如果目标模型能将其正确分类，则该攻击就认为这个样本是成员，如式(5)所示：

$$\mathcal{A}_{\text{corr}}(f,(\boldsymbol{x},y))=I\{\arg\max_i f(\boldsymbol{x})_i=y\} \tag{5}$$

其中，$I\{\cdots\}$是一个指示函数。该攻击的原理是：在模型过拟合的场景下，目标模型往往在训练集上具有较高的准确性，而在测试集上准确性相对较低。因此如果一个样本被正确预测，则从统计学的角度，该样本有更大的概率是成员。如果攻击者需要推断的样本点是成员和非成员的概率相同（都为 0.5），则该攻击的推断准确率可以由式(6)计算。

$$Adv(f)=1/2+(A_{\text{train}}-A_{\text{test}})/2 \tag{6}$$

其中，A_{train}和 A_{test}分别代表目标模型在训练集和测试集上的准确度。这个攻击方法较为简单，但是推断准确度相对较低，其性能一般作为其他方案比较的基准值。

基于 loss 的攻击利用目标模型在训练集和测试集上的 loss 差异。在基于 loss 的攻击中，攻击者首先训练一个影子模型来模仿目标模型，然后收集影子模型训练数据的平均 loss 值 δ 作为攻击的阈值。对于任意给定的样本点，如果模型对其的预测 loss 值小于设定的 loss 阈值 δ，则认为这个样本点是一个成员，计算过程如式(7)所示：

$$\mathcal{A}_{\text{corr}}(f,(\boldsymbol{x},y))=I\{\mathcal{L}(f(\boldsymbol{x}),y)<\delta\} \tag{7}$$

这个攻击的原理是：机器学习在训练的过程中，通过最小化模型在训练集上的预测 loss 值来更新模型的参数。因此，模型在训练集上的 loss 值应小于在其他数据上的预测 loss 值。

4.3 NSH 攻击

Nasr 等[28]首次在白盒场景下探究了成员推断问题。在白盒场景下，攻击者能获得目标模型的完全访问权限。对于任何样本点，攻击者不仅能获得目标模型对该样本的输出向量，还能获得模型在计算过程中产生的中间值（如神经网络隐藏层的计算值），因此可以收集更多信息以辅助成员推断过程。类似于 Shokri 等[20]的攻击，Nasr 等同样训练一个神经网络作为攻击模型。不同的是，对于每个样本点，灰盒的攻击模型只能获得目标模型对该样本的输出，Nasr 等提出使用样本点loss值相对于模型参数的梯度$\frac{\partial \mathcal{L}}{\partial \theta}$、目标模型各个隐藏层的计算值 $h_i(\boldsymbol{x})$和模型输出 $\theta(\boldsymbol{x})$等多方面特征作为攻击模型的输入来执行成员推断。

NSH 攻击作为白盒攻击，能获取更多有关目标模型的信息，因此推断准确率高于黑盒和灰盒攻击。但是在实际使用中，模型的梯度信息、loss 值等通常不会公布给普通使用者，因此白盒攻击通常作为模型拥有者检测自身模型隐私风险的工具。模型拥有者可以通过 NSH 攻击来系统地度量模型的隐私泄露风险。

4.4 ML-Leaks 攻击

Shorki 等所提出的 Shadow Training 攻击需要训练一个和目标模型相近的影子模型，因此需要拥有目标模型的训练知识、模型知识以及数据知识。Salem 等[21]提出了 ML-Leaks 攻击，包括 3 种模式，分别称为 Adversary I/II/III，旨在逐步放宽以上这些知识要求。

Adversary I 类似于 Shokri 的 Shadow Training 攻击，需要训练影子模型来模拟目标模型的行为，不同的是其攻击模型只取模型输出向量的 Top-3 置信度作为输入。实验证明，成员推断攻击不需要获取全部的置信度信息，部分（Top-3）输出就足够了。Adversary I 仍然需要模型知识、训练知识以及数据知识来训练一个高质量的影子模型。

Adversary II 改进了 Shokri 训练影子模型的方案，放宽了攻击需要知道的知识。他们假设攻击者不知道目标模型的类型，因此无法确定该训练哪种影子模型。在他们的方案中，攻击者训练多种影子模型，包括逻辑回归、决策树、神经网络等，并用这些模型产生带有成员标签的数据集来训练攻击模型。Salem 等声称，只要这些影子模型中包含目标模型的类型，就能实施有效的成员推断，因此不需要知道目标模型的训练知识和模型知识。

Adversary III 使用模型输出的最大预测向量来区分成员和非成员。具体来说，他们经观察发现，模型对于成员和非成员输出的预测向量中，成员的预测向量倾向于拥有更大的最大置信度。基于这一点，他们首先自己生成一部分样本点作为非成员，然后用这些样本点中最大的置信度作为阈值。对于一个样本，如果模型输出的置信度高于这个阈值，则认为该样本是成员。

4.5 BlindMI 攻击

具有高推断准确度的成员推断攻击大多需要使用 Shokri 等提出的 Shadow Training 方法训练一个影子模型，从影子模型获得参考来对目标模型执行成员推断。然而这种方法需要更多知识，难以在黑盒场景下应用。而其他不需要 Shadow Training 的攻击，攻击性能往往低于需要训练影子模型的方法。

Hui 等[29]针对这个问题，提出了一种不需要影子模型的攻击，称为 BlindMI。他们的攻击通过差分对比的方法直接分析目标模型的行为来执行成员推断。攻击的思想是：对于一个包含成员和非成员的数据集 A 和一个仅包含非成员的数据集 B，如果从 A 中移除一个点后，A 和 B 的整体距离更近了，则这个被移除的点是成员；反之是非成员。这里判断

两个数据集距离的方法是计算目标模型对两个数据集输出置信分布的差异。他们首先生成一批包含非成员数据点的数据集,然后将依次从需要推断的数据集(包含成员和非成员)中移除样本点并判断样本点是成员还是非成员。

4.6 Label-only 攻击

如第 3.1 节介绍,大部分成员推断攻击需要获得目标模型的输出知识,即能通过访问模型得到输出,但是模型返回的输出完整度有所不同。在现实场景中,很多部署的模型只返回硬标签(属于哪一个类别),而不返回具体的输出向量(各个类的置信度)。针对这个场景,Choquette-Choo 等[26]提出了 Label-only 攻击,仅通过目标模型返回的标签实施成员推断。在只能获取硬标签的情况下,Label-only 攻击通过观察一个样本在被目标模型分类过程中受干扰的难易程度来判断它是否是成员。例如,针对一个执行图像分类任务的目标模型,攻击者对一张图像样本不断添加噪声(如随机高斯噪声和对抗样本噪声),直到目标模型将这张图像分到另一个类中(硬标签改变)。攻击者根据扰动目标样本错误分类的难易程度,计算该样本到目标模型决策边界的距离 $dist_f(\boldsymbol{x},y)$,如果 $dist_f(\boldsymbol{x},y)>\tau$,则认为该样本是成员,其中 τ 是攻击者通过训练影子模型而预先设置的阈值。Choquette-Choo 等声称这种攻击的原理是样本抗扰动的鲁棒性可以一定程度上反映出该样本的置信度,从而在目标模型仅返回硬标签的情况下,也能从侧面获取到输出的置信度信息来进行成员推断攻击。

5 成员推断防御进展

成员推断攻击作为一种新型的针对机器学习的隐私攻击,揭示了机器学习存在的隐私风险,也引发了研究者对其防御方法的思考与研究。本节将简要总结代表性的成员推断防御方法,来描述目前成员推断防御领域的现状。

5.1 差分隐私

差分隐私[30]旨在从理论层面保证一个数据集中单个样本的隐私。具体来说,假设有两个数据集 D 和 D',且两个数据集仅有一个样本点 z 不同,对数据执行访问操作 M 以后,如果获得相同输出 O 的概率小于等于 e^{ε},则称 M 是 ε-差分隐私的,计算如式(8)所示:

$$Pr\{M(D)=O\}\leqslant e^{\varepsilon}\cdot Pr\{M(D')=O\} \tag{8}$$

Sablayrolles 等[31]指出,要保护机器学习训练中的样本点的成员隐私,则当模型在两个仅有一个样本不同的数据集 $D=\{z_1,\cdots,z_i,\cdots,z_n\}$ 和 $D'=\{z_1,\cdots,z_i',\cdots,z_n\}$ 上进行训练时,获得相同参数 θ 的概率应该符合式(9)。

$$\log\left|\frac{Pr(\theta|D)}{Pr(\theta|D')}\right|=|\mathscr{L}(\theta,z_i)-\mathscr{L}(\theta,z_i')| \tag{9}$$

其中,$\mathscr{L}$ 是模型的损失函数。

Abadi 等[32]提出了差分隐私思想在机器学习训练中的具体实现 DP-SGD。在训练中,DP-SGD 把每个样本的梯度剪切到 τ 的 L2 范式,并给每个 batch 的梯度添加高斯分布的扰动 $\mathcal{N}(0,c^2\tau^2)$,其中 c 和 τ 是预设的参数。DP-SGD 旨在通过添加扰动的方式来减小各个样本对模型的影响,从而保护每个训练样本的隐私。然而在实践中,Jayaraman 等发现,成功地保护隐私往往需要在训练梯度中添加大量扰动,从而导致模型性能明显下降。差分隐私能给保护成员隐私提供理论上的边界,但是在实际使用中仍需进一步优化。

5.2 对抗训练

Nasr 等[33]指出,传统的机器学习通常只考虑如何提升模型的性能,而没有在训练过程中考虑到模型被成员推断攻击的可能,因此容易泄露隐私。对此,他们提出了对抗训练方法,自己模拟攻击者来训练一个攻击模型,并在机器学习模型训练中增加攻击进行对抗训练来使模型同时最小化两个目标:1)在训练集上的分类错误;2)攻击模型的成员推断准确率。对抗性训练通过交替训练目标模型和攻击模型来不断提升防御的效果。

具体来说,防御者的目标是训练一个受保护的目标模型 θ,使得 θ 能抵御成员推断攻击。为了实现这个目标,Nasr 等需要同时训练一个攻击模型 A 来实施成员推断攻击。在每个训练轮次中,首先根据式(10)对攻击模型 A 进行更新,使得攻击模型在目标模型训练集(成员)和测试集(非成员)上的推断准确率最大化。

$$\arg\max_{A}\mathbb{E}_{(x,y)\sim\mathbf{Pr}_{D_{\mathrm{tr}}}}[\log A(\theta(\boldsymbol{x}),y)]+\mathbb{E}_{(x,y)\sim\mathbf{Pr}_{\backslash D_{\mathrm{tr}}}}[1-\log A(\theta(\boldsymbol{x}),y)] \tag{10}$$

随后根据公式更新目标模型 θ,同时最小化目标模型的分类损失和被攻击的可能性。

$$\arg\min_{\theta}\sum_{(x,y)\in D_{\mathrm{tr}}}\mathscr{L}(f(\boldsymbol{x}),y)+\lambda\log(A(f(\boldsymbol{x}),y)) \tag{11}$$

其中,λ 是一个超参数,控制目标模型性能和隐私的权衡。使用对抗训练方法训练目标模型时,相当于在目标函数上增加了一个正则化项,将攻击模型的攻击效果考虑在内,约束目标模型的优化,从而达到防御攻击的目的。对抗训练首次引入了新的训练框架来防御成员推断攻击,但是 Song 等[34]在实验中发现,使用对抗性训练难以训练出既能防御攻击又保持高性能的模型。

5.3 Memguard

差分隐私和对抗训练等防御方法通过影响模型的训练过程来达到防御的效果。这些方法在抵御成员推断的同时,往往会导致模型的性能出现明显下降。Jia 等[35]针对这个问题提出了一种黑盒的防御,称为 Memguard。Memguard 不影响模型的训练,而是通过对模型的输出添加扰动来迷惑攻击者。

Jia 等分析了黑盒成员推断攻击的形式,发现很多成员推断攻击需要攻击者训练一个神经网络模型作为攻击模型来推断成员。而神经网络被证明容易受到对抗样本攻击[36-38]:只需要对正常样本添加细微的扰动,就可以欺骗神经网络,使其错误分类添加扰动后的样本。Jia 等利用神经网络的这个弱点,设计了 Memguard。给定一个目标模型 θ,Memguard 对模型输出 $\theta(\boldsymbol{x})$ 添加噪声来欺骗攻击者的攻击模型 A,使得 A 无法正确地将对应的输出区分为成员或者非成员。

攻击模型 A 以每个样本的预测向量 $\theta(\boldsymbol{x})$ 为输入,并输出该样本是成员的概率 $A(\theta(\boldsymbol{x}))$。和对抗性训练一样,Memguard 需要防御者自己训练一个攻击模型来参与防御过程。Memguard 防御的关键是如何对 $\theta(\boldsymbol{x})$ 添加扰动 n,使 A 错误分类 $\theta(\boldsymbol{x})$ 且添加扰动后的模型输出不影响正常的使用。Jia

等将这个目标定义为了以下问题进行求解。

$$\begin{aligned}&\min_{n} d(\theta(\boldsymbol{x}),\theta(\boldsymbol{x})+n)\\&\text{s.t.}\arg\max_{i}(\theta(\boldsymbol{x})_i+n_i)=\arg\max_{i}(\theta(\boldsymbol{x}_i))\\&\quad A(\theta(\boldsymbol{x})+n))=0.5\\&\quad \theta(\boldsymbol{x})_i+n_i\geqslant 0\\&\quad \forall i\sum_i n_i=0\end{aligned}\tag{12}$$

其中，求解的目标是最小化原始预测向量和加扰动后的向量之间的距离 d。需要满足的限定条件包括：添加扰动后分类结果不变；攻击模型预测该加扰样本为成员的概率接近 0.5；加扰后的置信度向量里，所有置信度都大于零，且和仍为 1。

Memguard 不影响模型的训练，也不改变模型输出的分类结果，而只改变置信度的值。因此，使用 Memguard 不会导致模型分类准确率的下降。但是 Memguard 产生的对抗样本只对使用神经网络的特定攻击有效，对基于 Metric 的攻击和 Label-only 攻击效果不佳。

5.4 知识蒸馏

知识蒸馏最早由 Hinton 等提出，用于提升参数量较小的模型的性能[39-40]。知识蒸馏使用一个大模型（也叫老师模型）的输出来训练一个小模型（也叫学生模型），将老师模型的“知识”转移给学生模型。给定一个训练集 D 和一个训练好的大模型 θ_t，学生模型 θ_s 的任务是拟合 θ_t 在 D 上的输出，学习老师模型在数据集 D 上的行为。

Shejwalkar 等[41]将知识蒸馏引入成员推断攻击的防御中。假设需要保护成员信息的数据集是 D_{pr}，他们首先使用 D_{pr} 训练一个不受保护的模型 θ_{up}；随后使用 θ_{up} 在另外一个数据集 D_{rf} 上的输出来训练一个受保护的模型 θ_p；最后将 θ_p 作为被最终发布的模型供用户使用。这种方法主要通过限制训练过程中 θ_p 对隐私数据集的 D_{pr} 的直接访问来保护 D_{pr} 的成员隐私。θ_p 并没有直接在 D_{pr} 上进行训练，而是通过知识蒸馏，模拟老师模型 θ_{up} 在另外一个数据集上的行为来获得性能，因此不会泄露 D_{pr} 中的隐私。Shejwalkar 等证明，基于知识蒸馏的防御不仅能将成员推断攻击的准确率降低到接近于随机猜测的程度，而且对模型的性能没有明显影响。知识蒸馏防御的缺点是需要一个额外的数据集，从而限制了其在数据量有限的场景下的应用。

6 成员推断攻击成因分析

成员推断攻击原理的分析是一个很有挑战的研究问题。研究人员通常从实证的角度，用实验现象分析各种因素对成员推断攻击的影响 。

Shokri 等[20]发现，目标模型的过拟合程度越大，模型越容易受到成员推断攻击。Yeom 等[22]分析了这个现象，并证明过拟合时，模型在训练集和在测试集上的表现不一致。例如，模型对训练集有更高的准确率和更低的 loss 值。他们用实验说明这些不一致性可以被攻击者用来有效地区分成员和非成员，并证明过拟合是成员推断攻击的充分条件。

除了过拟合以外，研究人员还发现目标模型的类别也是影响成员推断攻击的重要因素。即使在相同的过拟合程度下，不同模型受成员推断攻击影响的程度也可能不同。Truex 等[42]发现，相较于朴素贝叶斯模型，决策树更容易被成员推断攻击。他们声称这是因为朴素贝叶斯算法独立地考虑每个特征对于给定类的概率，而给定大量的训练样本时，单个样本对这些概率的影响很小。相反，决策树通过考虑特征之间的组合情况来给出分类结果，而不是独立看待这些特征，因此受单个样本影响更大。假如某个样本具有独特的特征组合形式，甚至可能导致决策树生成一个新的分支。因此，相比贝叶斯，决策树的决策边界更容易受单个样本影响，使得攻击者能容易区分每个样本是否参与了模型的训练。

最后，研究人员发现模型受成员推断攻击的程度也和训练集数据的特点有关。Salem 等[21]的实验发现，训练集数据的类别越多，模型越容易受到成员推断攻击。针对这个现象，Turex 等[42]分析认为，对于一个相同的输入空间，更多的类别意味着每个类别的特征空间区域更小。这将使任何单个样本更有可能改变决策边界，因为区域之间的空间“更紧密”。Turex 等[43]还发现，当模型的训练集中样本特征的方差较小时，模型更难被攻击。对于每个类中的样本来说，特征的方差小代表一个实例与同一类的其他实例相似。在这种情况下，单一样本不太可能显著影响模型的决策边界，因此攻击者更难推断成员。

总的来说，成员推断攻击与目标模型过拟合程度、目标模型算法类别和训练集特征等多方面因素有关。相比快速发展的成员推断攻防技术，对成员推断攻击原理的研究仍处于萌芽阶段，有待进一步探索。

7 挑战与展望

作为一个新兴的研究领域，成员推断攻击暴露出了机器学习中存在的隐私问题，引发了学界和工业界的广泛关注。目前针对成员推断的研究主要集中在攻击、防御和原理揭示 3 个方面。在攻击方面，研究人员不断地探究更加具有威胁的攻击方案，来尽可能暴露更多的隐私漏洞，从而给后续防御方案的研究提供参考。例如，Shokri 等[20]最早提出的第一个成员推断攻击需要知道目标模型的结构、训练知识等信息，而最近 Hui 等[29]提出的 BlindMI 攻击不需要知道任何有关目标模型的知识就能实施高准确度的成员推断。在防御方面，研究人员正在根据成员推断攻击的基本原理探究相应的对策，目前的挑战主要集中在隐私和模型性能的权衡上。例如，差分隐私[32]、对抗训练[33]等防御能有效降低成员推断的效果，但是对模型性能的影响较大；而 Memguard 等不影响训练的防御方法只对特定攻击有效，难以进一步拓展[35]。如何在达到良好防御效果的同时，尽量减小对模型可用性的影响，是一个重要的研究方向。另一方面，研究人员也在不断探究成员推断攻击产生的原因，从原理层面解释成员推断背后的原因，为攻击和防御的研究提供参考。

总的来说，成员推断攻击的研究对机器学习中的隐私保护有重要的指导意义，其未来的发展方向至少包括 3 个方面：1）攻击方可以在在更严格的威胁模型下，研究成员推断攻击的可行性，例如研究无需影子模型或者无需输出向量的成员推断攻击，从而更全面地揭示其隐私威胁；2）防御方可以考虑结合机器学习其他领域的技术，如对抗性网络[43]、知识

蒸馏[44]等，针对成员推断攻击设计出更加实用的防护方案，做好隐私和性能的权衡；3)探究机器学习中影响成员推断攻击效果的因素，如模型过拟合程度[22]、数据集分布[42]等，揭示成员推断攻击的原理。

结束语 随着机器学习技术的快速发展，机器学习模型被应用于包括人脸数据、医疗数据在内的多种隐私数据的处理，而随之而来的隐私泄露风险也引发了人们的广泛关注。成员推断攻击作为一种直接针对机器学习中隐私数据的推断攻击，其研究对于保护机器学习的隐私具有重大意义。本文介绍了针对机器学习的成员推断攻击的基本原理和分类方法；重点选取了6种代表性攻击方案进行了详细阐述和分析，尤其是从攻击方所需条件的角度解读了成员推断攻击在发展过程中的主要进展；随后介绍了成员推断的4种主流防御手段，并仔细分析了各方案的优缺点；最后回顾成员推断的发展历程，针对现有攻击和防御方法的不足，指出了目前研究面临的挑战，并给出了该领域未来的3个重要研究方向。

参 考 文 献

[1] WEYAND T, ARAUJO A, CAO B Y, et al. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020: 2572-2581.

[2] HENAFF O. Data-efficient image recognition with contrastive predictive coding[C]// 2020 International Conference on Machine Learning(ICML). 2020: 4182-4192.

[3] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv: 2005. 14165, 2020.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-418.

[5] DOUMBOUYA M, EINSTEIN L, PIECH C. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users[C]// 2021 AAAI Conference on Artificial Intelligence(AAAI). 2021: 14757-14765.

[6] LIU S, GENG M, HU S, et al. Recent progress in the cuhk dysarthric speech recognition system[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2267-2281.

[7] TAIGMAN Y, YANG M, RANZATO M, et al. Deepface: Closing the gap to human-level performance in face verification [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2014: 1701-1708.

[8] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]// 2015 IEEE Conference on Computer Vision and Patten Recognition (CVPR). 2015: 815-823.

[9] ERICKSON B J, KORFIATIS P, AKKUS Z, et al. Machine learning for medical imaging[J]. RadioGraphics, 2017, 37(2): 505-515.

[10] KOUROU K, EXARCHOS T P, EXARCHOS K P, et al. Machine learning applications in cancer prognosis and prediction [J]. Computational and Structural Biotechnology Journal, 2015, 13: 8-17.

[11] CARLINI N, LIU C, ERLINGSSON Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks[C]// 28th USENIX Security Symposium(USENIX Security 19). 2019: 267-284.

[12] SONG C, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security(CCS 17). 2017: 587-601.

[13] LEINO K, FREDRIKSON M. Stolen memories: Leveraging model memorization for calibrated white-box membership inference[C]// 29th USENIX Security Symposium(USENIX Security 20). 2020: 1605-1622.

[14] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[C]// 25th USENIX Security Symposium(USENIX Security 16). 2016: 601-618.

[15] OH S J, AUGUSTIN M, FRITZ M, et al. Towards reverse-engineering black-box neural networks [C] // 2018 International Conference on Learning Representations(ICLR). 2018: 1-20.

[16] YU H, YANG K, ZHANG T, et al. Cloudleak: Large-scale deep learning models stealing through adversarial examples [C] // 2020 Network and Distributed System Security Symposium (NDSS). 2020: 1-16.

[17] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// 2015 ACMSIGSAC Conference on Computer and Communications Security(CCS 15). 2015: 1322-1333.

[18] ZHANG Y, JIA R, PEI H, et al. The secret revealer: Generative model-inversion attacks against deep neural networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020: 250-258.

[19] MEHNAZ S, LI N, BERTINO E. Black-box model inversion attribute inference attacks on classification models [J]. arXiv: 2012. 03404, 2020.

[20] SHOKRI R, STRONATI M, SONG C, et al. Membership Inference Attacks against Machine Learning Models [C] // 2017 IEEE Symposium on Security and Privacy(SP). 2017: 3-18.

[21] SALEM A, ZHANG Y, HUMBERT M, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models[C]// 2019 Network and Distributed Systems Security(NDSS) Symposium. 2019: 1-15.

[22] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting [C]// 2018 IEEE 31st Computer Security Foundations Symposium(CSF). 2018: 268-282.

[23] GERUM R C, ERPENBECK A, KRAUSS P, et al. Sparsity through evolutionary pruning prevents neuronal networks from overfitting[J]. Neural Networks, 2020, 128: 305-312.

[24] SONG X, JIANG Y, TU S, et al. Observational overfitting in reinforcement learning [C] // 2020 International Conference on Learning Representations(ICLR). 2020: 1-29.

[25] RICE L, WONG E, KOLTER Z. Overfitting in adversarially ro-

bust deep learning[C]// The 37th International Conference on Machine Learning(ICML). 2020:8093-8104.

[26] CHOQUETTE-CHOOCA C A,TRAMER F,CARLINI N, et al. Label-only membership inference attacks[C]// The 38th International Conference on Machine Learning(ICML). 2021: 1964-1974.

[27] KRIZHEVSKY A,HINTON G,et al. Learning multiple layers offeatures from tiny images[R]. Technical report,University of Toronto,2009.

[28] NASR M,SHOKRI R,HOUMANSADR A. Comprehensive privacy analysis of deep learning:Passive and active white-box inference attacks against centralized and federated learning[C]// 2019 IEEE Symposium on Security and Privacy(SP). 2019:739-753.

[29] HUI B,YANG Y,YUAN H,et al. Practical blind membership inference attack via differential comparisons[C]// Network and Distributed Systems Security(NDSS) Symposium. 2019:1-17.

[30] DWORK C,MCSHERRY F,NISSIM K,et al. Calibrating noise to sensitivity in private data analysis[J]. Journal of Privacy and Confidentiality,2017,7(3):17-51.

[31] SABLAYROLLES A,DOUZE M,SCHMID C,et al. White-box vs black-box:Bayes optimal strategies for membership inference [C]// The 36th International Conference on Machine Learning (ICML). 2019:5558-5567.

[32] ABADI M,CHU A,GOOD-FELLOW I,et al. Deep learning with differential privacy[C]// 2016 ACMSIGSAC Conference on Computer and Communications Security(CCS 16). 2016: 308-318.

[33] NASR M,SHOKRI R,HOUMANSADR A. Machine learning with membership privacy using adversarial regularization[C]// 2018 ACM SIGSAC Conference on Computer and Communications Security(CCS 18). 2018:634-646.

[34] SONG L W,MITTAL P. Systematic evaluation of privacy risks of machine learning models[C]// 30th USENIX Security Symposium(USENIX Security 21). 2021:2615-2632.

[35] JIA J,SALEM A,BACKES M,et al. Memguard:Defending against black-box membership inference attacks via adversarial examples[C]// 2019 ACM SIGSAC Conference on Computer and Communications Security(CCS 19). 2019:259-274.

[36] PAPERNOT N,MCDANIEL P,GOODFELLOW I,et al. Practical black-box attacks against machine learning[C]// 2017 ACM on Asia Conference on Computer and Communications Security(AsiaCCS 17). 2017:506-519.

[37] CARLINI N,WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy(SP). 2017:39-57.

[38] TRAMèR F,KURAKIN A,PAPERNOT N,et al. Ensemble adversarial training:Attacks and defenses[C]// 2018 International Conference on Learning Representations(ICLR). 2018:1-20.

[39] HINTON G,VINYALS O,DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503. 02531,2015.

[40] DU S,YOU S,LI X,et al. Agree to disagree:Adaptive ensemble knowledge distillation in gradient space[C]// 2020 Advances in Neural Information Processing Systems(NeurIPS 20). 2020: 12345-12355.

[41] SHEJWALKAR V,HOUMANSADR A. Membership privacy for machine learning models through knowledge transfer[C]// 2021 AAAI Conference on Artificial Intelligence. 2021: 9549-9557.

[42] TRUEX S,LIU L,GURSOY M,et al. Demystifying membership inference attacks in machine learning as a service[J]. IEEE Transactions on Services Computing,2021,14(6):2073-2089.

[43] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[C]// 2014 Advances in Neural Information Processing Systems(NeurIPS 14). 2014:1-9.

[44] CHO J H,HARIHARAN B. On the efficacy of knowledge distillation[C]// 2019 IEEE/CVF International Conference on Computer Vision(ICCV). 2019:4793-4801.

PENG Yuefeng, born in 1998, postgraduate. His main research interests include artificial intelligence security and so on.

ZHAO Bo, born in 1972, Ph.D, professor, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include trusted computing and trustworthy artificial intelligence.

(责任编辑:柯颖)