机器学习中的隐私保护研究和前瞻

黄磊 计1802 20184777

April 30, 2021

Abstract

新出现的机器学习和深度学习方法已经成为一股强大的推动力，推动了智能医疗保健、金融技术和监控系统等广泛行业的变革。与此同时，在这个基于机器学习的人工智能时代，隐私已成为一个大问题。值得注意的是，机器学习环境下的隐私保护问题与传统数据隐私保护中存在很大的不同，因为机器学习可以是朋友也可以是敌人。目前，关于隐私保护和机器学习的工作还处于起步阶段，现有的解决方案大多只关注机器学习过程中的隐私问题。因此，需要对隐私保护问题和机器学习进行全面的研究。本文综述了机器学习中隐私问题和解决方案的最新进展。本文涵盖了隐私与机器学习交互的三种类型:(i)私人机器学习，(ii)机器学习辅助隐私保护，(iii)基于机器学习的隐私攻击及其保护方案。对每一类的研究进展进行了回顾，并指出了主要的挑战。最后，基于对隐私和机器学习领域的深入分析，指出了该领域未来的研究方向。

关键词：信息安全；机器学习；隐私安全

1 介绍

自2018年Facebook数据隐私丑闻以来，隐私安全再次引起了人们的重视。这促使人们重新审视隐私问题，特别是在大数据革命的推动下，随着智能技术的出现，例如，新兴的机器学习技术和深度学习技术将对隐私保护产生巨大的影响。需要深入研究的一个关键问题是：与机器学习相关的隐私挑战和解决方案是什么?

机器学习和隐私保护相关的学术研究也逐渐开始涌现，其中多数是关于机器学习模型相关的隐私挑战和风险，强调在机器学习过程中降低隐私风险。在这方面，提出了可能的攻击模型 [9, 40, 3, 36, 25, 37]，相应的防护方案 [1, 35, 5, 30]也已经被提出。这些工作表明，机器学习模型和训练数据集都可以成为隐私攻击的目标，导致敏感信息泄漏。与此同时，研究人员也试图利用机器学习来保护隐私。例如，[41]的作者开发了一种自动识别隐私敏感对象类并调整用户隐私偏好设置的方法。此外，也有一些学者在基于机器学习攻击模型的背景下提出了更新的隐私保护方案 [21, 20]。总的来说，目前的研究只是触及了表面，还有一些主要问题需要进一步研究:

- 机器学习可以在隐私保护问题中扮演不同的角色，例如，保护目标、攻击工具或保护工具。它甚至可能在同一个问题中扮演多个角色。
- 机器学习系统和模型有不同的类型，每个类型面临不同的隐私风险，需要不同的保护方案。
- 没有一个统一的隐私标准或概念。尽管差分隐私(Differential Privacy) [8]在传统的隐私研究中被广泛接受，但它在机器学习背景下仍有局限性，特别是在考虑非结构化数据时，如文本、图像和视频。

在这种情况下，隐私保护和机器学习的研究十分必要。虽然有一些关于这个主题的研究[14, 2, 42, 23]，但焦点一直在某种类型的机器学习模型或特定的方法上。本文通过研究隐私和机器学习的不同场景和应用，对机器学习中的隐私问题进行全面的研究。本文的主要内容如下:

- 按照机器学习扮演的不同角色，即机器学习作为保护目标(私有机器学习)、保护工具(机器学习增强隐私保护)、攻击工具(基于机器学习的攻击)，对这一领域的工作进行了划分，并分析了每种类型中存在的问题和解决方案。
- 对于私有机器学习，对攻击和保护方案进行分类，然后比较它们的区别。
- 对于基于机器学习的隐私保护和基于机器学习的隐私攻击，不仅讨论了现有的研究工作，也提供了实现隐私保护的新技术的见解。
- 研究最后讨论了未来机器学习和隐私研究的方向。

本文的其余部分组织如下。第2节回顾了机器学习系统和模型的基本概念，并讨论了隐私和机器学习之间的关系。第3节讨论了基于机器学习的攻击和相应的隐私保护方案。在第4节中，提出了对未来机器学习和隐私保护研究的展望，并提出了一些未来的研究方向。最后，在第5节中通过一个总结来结束该文的工作。此外，本文中使用的缩写如表 1所示。

Table 1: Summary of acronyms used in the paper.

CNN	convolutional neural network	卷积神经网络
DNN	deep neural network	深度神经网络
DP*	differential privacy	差分隐私
ERM	empirical risk minimization	经验风险最小化
FGSM	fast gradient sign method	快速梯度符号方法
FHE	fully homomorphic encryption	同态加密
GAN	generative adversarial network	生成对抗网络
GNN	generative neural network	图神经网络
IoT	Internet of things	物联网
ML	machine learning	机器学习
SGD	stochastic gradient descent	随机梯度下降
SMC	secure multi-party computation	安全多方计算
SVM	support vector machine	支持向量机
VAE	variational autoencoder	变分自编码器

2 隐私威胁与机器学习

在本节中，本文将讨论机器学习背景下的隐私威胁，并进一步指出机器学习在用户隐私研究中的各种作用。

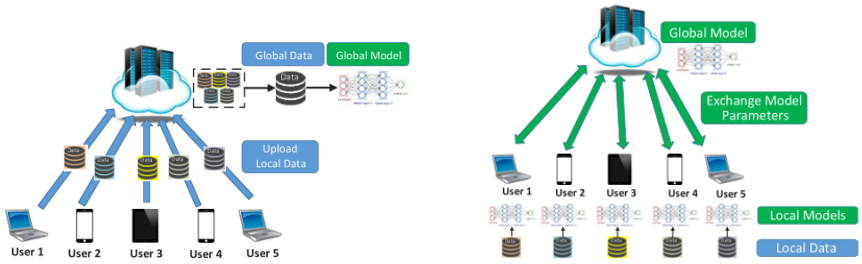
2.1 机器学习系统和模型

机器学习指的是计算机系统使用的算法和统计模型，可以在不使用明确指令的情况下高效地执行特定任务。它依赖于自动学习过程。机器学习算法为样本数据构建了一个数学模型，称为“训练集”，用来做出预测或决策 [4].

根据输出是否在训练集中被标记，机器学习模型可以分为三组:监督、无监督和半监督。由于监督学习被大多数实际的机器学习算法所使用，这里将其作为一个典型例子进行解释。

监督机器学习模型是一个参数化函数 f_θ ，是输入数据 $\vec{x} \in \mathbb{X}^d$ (通常是特征向量)到输出数据的映射 $y \in \mathbb{Y}$ (标签). 对于一个分类问题 $\mathbb{X}^d$ 是一个 d-维的向量空间而 $\mathbb{Y}$ 是类别的集合. 其功能被训练成准确预测以前没有见过的新数据的标签。

此外，我们可以将机器学习过程分为两个阶段:



(a) centralized learning (can be outsourced learning).

(b) distributed learning

Figure 1: Centralized and distributed ML systems: (a) centralized learning; (b) distributed learning.

1. 模型的训练阶段: 一个机器学习模型的训练过程是寻找能够精确捕捉 $\mathbb{X}$ and $\mathbb{Y}$之间关系的最优参数. 为了实现这个目标，需要一个大小为N的训练集$D = \{\vec{x}_i, y_i\}_{i=1}^{N}$ 然后采用一个损失函数L来量化两个输出之间的差异,即实际值和预测值$f_\theta(\vec{x}_i)$. 训练模型的目标是最小化这个损失函数，

$$\theta^\star = \arg\min_\theta (\sum_i L(y_i, f_\theta(\vec{x}_i)) + \Omega(\theta)), \tag{1}$$

   其中Ω是一个正则化项,用于惩罚模型的复杂性和避免过拟合。

2. 模型的预测阶段: 在模型训练完成并得到最优参数 $\theta^\star$ 之后，给定输入$\vec{x}$，则相应的输出可以计算为$y = f_\theta^\star(\vec{x}_i)$。这个预测过程叫做推断。我们可以在测试数据集$D_T$上计算模型的预测精度来衡量模型的性能。

此外，根据机器学习系统的架构，有两种不同的模型，如图1所示:

- 集中学习:将训练数据集中在一台机器或一个数据中心中，由集中的实体对模型进行训练和托管。例如，研究人员可以使用云平台托管数据集，并在此基础上训练人工智能模型。毫无疑问，在这种集中的方法中，所有数据的可用性导致了 [12]的高效率和准确性。但是，由于集中式运营商直接访问敏感数据，可能会侵犯用户隐私。

  随着学习任务变得越来越复杂，许多公司开始外包训练过程，也叫*outsourced learning*, or *ML-as-a-service*. 在这种情况下，每个用户拥有自己的训练数据，而服务提供者拥有模型和算法。数据持有者将模型创建外包给像Microsoft Azure ML和Amazon AWS ML这样的云服务，这些云服务可以自动化机器学习的过程。“用户上传数据集，进行训练，并使生成的模型可用” [38]。在此过程中，用户对模型创建的细节没有任何了解。“机器学习提供者是向数据持有者提供机器学习训练代码的实体” [38]。

- 分布式学习:集中学习有时不是一个好的选择，原因如下:(i)数据在某些场景中本质上是分布式的;(ii)数据太大，无法储存在一台机器上;(iii)用户不愿意分享原始数据;(iv)用户希望通过不样本训练神经网络来获得更好的预测精度。在这种情况下，机器学习可以以分布式的方式进行，即分布式学习。分布式学习一般用于分布式训练数据源和集中式服务器的场景。分布式学习有几种变体:

  – 协作学习:涉及这种协作的分布式学习被称为协作学习。但是相关研究中的场景可能会大不相同。例如， [37]的作者提出了一个协作学习框架，该框架“同时在同一训练数据上”训练多个分类器，以获得更好的性能。另一方面，在 [12]定义的协作学习模型中，每个参与者使用自己的设备来训练本地的AI模型。然后，它与其他用户共享模型的一部分参数。服务运营商可以通过收集这些参数来创建复合模型，并获得与使用集中式方法构建的模型几乎相同的准确性。

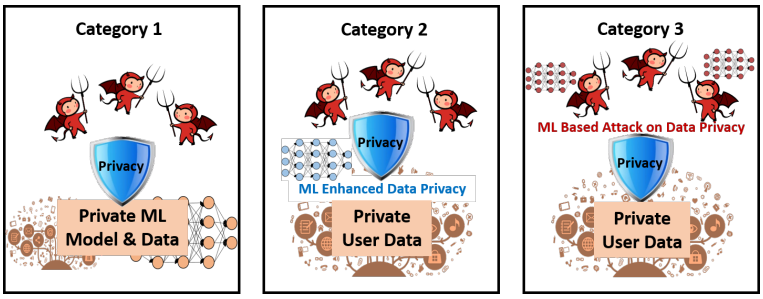


Figure 2: Three different categories of research problems in privacy and ML: (a) Privacy of ML model and data; (b) ML enhanced privacy protection; (c) ML-based privacy attack.

协作的方法“更加隐私友好”，因为数据集不是直接暴露的。同时,如果模型中只有一小部分参数是共享的,并且参数被DP机制截断和混淆,则通过实验证明模型具有收敛性 [35].

– 联邦学习:一个流行的协作学习框架是由谷歌引入的联邦学习 [16]。目前有两种不同的联邦学习设置:跨设备和孤井互联 [15]。跨设备设置通常涉及非常多的移动或物联网设备，而在孤井互联设置中，它“可能只涉及少量相对可靠的用户主机” [15]，例如多个组织。在涵盖这两种设置的更广泛的联邦学习定义中，每个设备从集中式服务器下载当前模型，通过学习本地设备上的数据来改进它，然后在有重点的更新中总结更改。在这里，“聚焦更新”是包含“特定学习任务所需的最低信息”的更新 [15]。然后通过平均所有用户的更新来更新共享模型。由于所有的训练数据都不会离开本地设备，而且个人用户的更新也不会存储在云中，因此隐私风险大大降低。

– 分割学习:另一个协作学习框架是分割学习，每个用户训练网络到一个特定的层，称为切割层，并将权重发送到服务器。从数学上讲，这些权值代表并压缩输入数据到一些中间特征向量。然后服务器对剩下的层进行训练，生成最后一层的梯度，然后进行误差反向传播，直到切割层。然后渐变被传递给用户。其余的反向传播由用户完成。在分割学习中，“由于要传输的数据被限制在分裂神经网络的前几层，客户端通信成本显著降低”。

尽管一些协作学习模型考虑共享训练数据，这带来了重大的隐私风险。然而，在本文中考虑了本地原始训练数据没有与服务器或用户之间共享的情况。在这个学习过程中，用户可以协作学习一个共享的机器学习模型，从而将机器学习任务与存储在单个设备中的数据解耦。

总体而言，集中式学习的特征是“全局存储数据”和“全局训练模型”，如图1(a)所示，而分布式学习的特征是“本地存储数据”和“本地训练模型”，如图1(b)所示。虽然分布式学习将会有一个全局性的模型，但它不是全局性的训练，至少模型的一部分是由单个用户训练的。

2.2 隐私与机器学习的联系

与传统的隐私相关研究框架相比，机器学习技术为隐私保护带来了新的挑战和机遇。已经有一些初步的研究开始了这方面的探索。根据机器学习在隐私方面的作用，现有的工作可以分为三类

第一，将机器学习系统私有化，即将机器学习系统作为隐私保护的目标。如图2(a)所示，类1包括将机器学习系统(模型参数)和数据(训练/测试数据集和输出数据)私有，因为隐私威胁可能发生在数据周期的任何阶段，例如数据的训练、发布或预测。该类的大部分研究依赖于在机器学习和深度学习模型中使用差异隐私。例如，Shokri等人 [35]开发了一种差分隐私SGD算法和分布式深度学习模型训练系统。通过这种方式，多个实体可以协同学习一个神经网络。

第二，使用机器学习增强隐私保护。如图2(b)所示，隐私保护目标是类2中的数据，机器学习是帮助隐私保护的工具。例如, Liu等人 [22]利用机器学习来增强决策隐私。Orekondy等 [28]提出了一种对图像

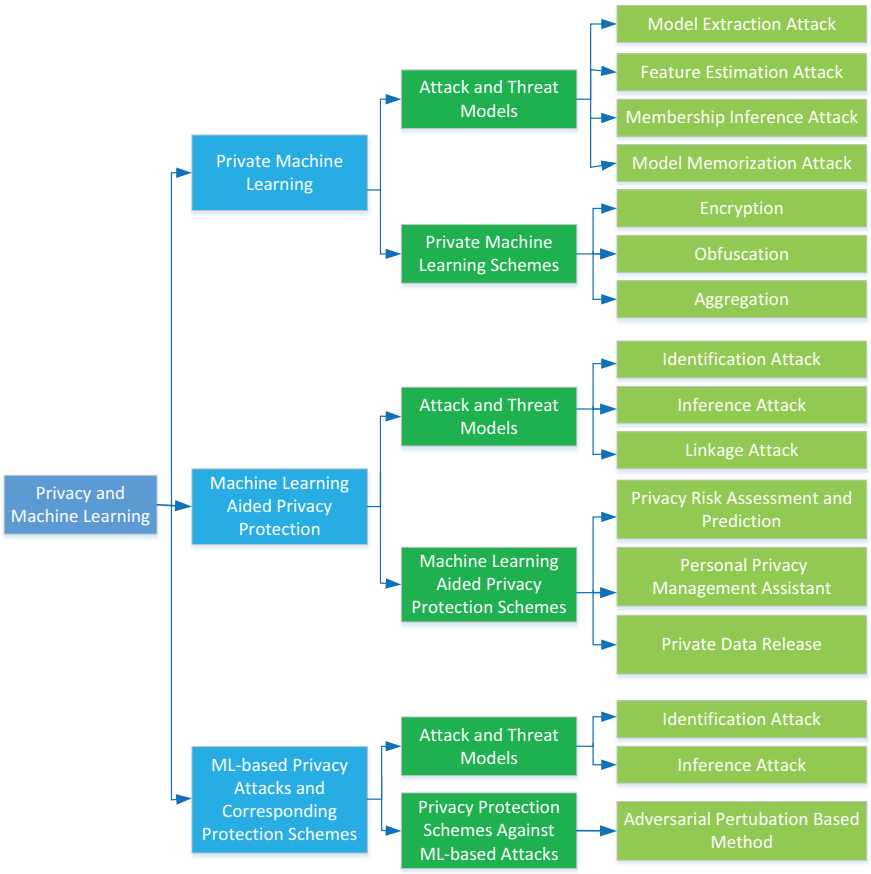


Figure 3: The proposed taxonomy of privacy and ML.

中的个人信息进行分类并直接预测图像中信息泄漏的方法。Yuan等人 [21]提出了一种机器学习方法来决定是否与特定上下文的特定请求者共享图片。

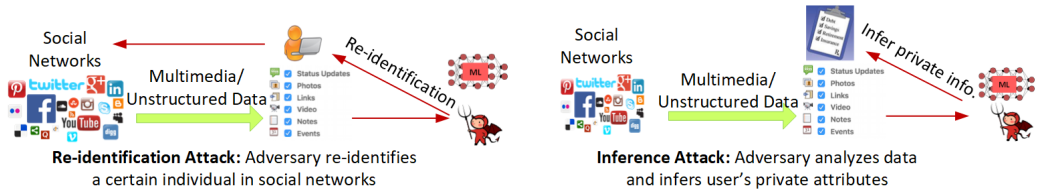
第三，基于机器学习的隐私攻击，即使用机器学习作为对手的攻击工具，如图2(c)所示。例如，最近的研究表明，深度学习方法可以用于从发布在互联网上的图像中检测物体类型、人的身份和地标。当对手使用这种强大的工具时，传统的隐私保护方法将被超越，特别是受到强大的深度学习工具的挑战。这类研究很少。Liu等人[20]提出了应用对抗性摄动图像的方案，从而使机器学习系统无法从中获得私有信息。

表2总结了涉及机器学习系统的三类隐私保护问题。值得一提的是，一种技术可能属于多个类别。例如，机器学习可能同时用作攻击和保护工具，这使问题更加复杂。本文将在接下来更详细地讨论这个问题。

Table 2: Three categories of privacy protection problems in the context of Machine Learning.

Category	**Role of ML in Privacy Protection**
Private ML	Protection target
ML enhanced Privacy Protection	Protection tool
ML-based Privacy Attack	Attack tool

图3总结了本文研究论文的一般分类。本文按照上述三类进行分类，在每个类别中，本文首先讨论了攻击和威胁模型，然后分析了隐私保护方案的工作



(a) Identification attack (b) Inference attack

Figure 4: Different privacy attack and threat models when ML is used as the attack tool.

3 基于机器学习的隐私攻击和相应的保护方案

机器学习既可以用作隐私保护工具，也可以用作攻击工具。它敦促我们重新审视隐私的定义和范围。特别是新兴的深度学习技术可以“自动收集和处理数百万张照片或视频，从社交网络中提取隐私或敏感信息” [20]。传统的隐私保护方法在与深度学习工具对抗时显得过于强大。接下来将讨论机器学习带来的新的隐私威胁和相应的解决方案。

3.1 攻击和威胁模型

个人信息泄露风险最大的来源是社交网络。虽然各种各样的社交网络平台丰富了人们的互动性和关系，但分享的帖子包括打卡、活动、想法(推文、状态更新等)、图片、视频等往往伴随着敏感信息。这些信息具有很高的隐私风险，他们有可能在无意中泄露自己的隐私。越来越多的公司和初创企业专门分析社交媒体上分享的照片，以利用它们实现商业目的，或将它们卖给其他公司。因此，最先进的DNN被用来进行隐私攻击。

例如，对手可以使用地理位置信息发起本地化攻击，主要包括定位个人的位置和时间信息。Gu等人和Mahmud等人 [11, 24]展示了一种危险的攻击，其目的是“定位关键的地点，如家庭和工作场所”。已经有一些研究讨论了家庭位置识别问题，要么基于“帖子内容”，要么基于“打卡”。“研究表明，在许多情况下，识别准确率可能超过90%” [21]。

多媒体数据除了具有简单的位置信息外，在机器学习工具的攻击下还具有更大的风险。公司利用先进的DNN群集照片来推断用户的偏好，以方便营销人员发送有针对性的广告。DNN被认为是机器学习中最实用的工具之一，因为它们利用了高效的训练算法和大数据集，这使它们能够比其他现有的机器学习技术表现得更好。这类机器学习工具的威力本身就成了一个问题，一旦照片被分享到社交媒体上，就可能会损害照片的隐私，也是一个需要解决的挑战性问题。

在物联网环境中，敏感数据、照片和视频的隐私变得更加重要，因为用户甚至可能不知道自己的信息(如照片和视频)正在被录制。例如，监控摄像头控制的区域可能会严重损害用户隐私，因为人们无法知晓自己的照片和视频是如何被捕捉和管理的。该监控系统很可能在未经用户许可的情况下应用人脸识别和检测等技术来识别用户。2014年皮尤互联网调查报告显示，超过91%的参与者“强烈同意”或“同意”“他们对自己的个人信息被公司收集和使用”的方式无法容忍。主要的机器学习攻击模型有再识别攻击和推理攻击，如图4所示。

- 再识别攻击可以通过人脸识别技术发起。DNN的最新进展使其从两个方面更加有害。首先，该过程实现了高精度的自动化。其次，采用模糊处理等传统保护方案不再有效。图4(a)是再识别攻击的一个例子。

- 当装备了机器学习时，推理攻击也变得更加强大。机器学习分类器可以用来从目标用户的公共数据(如twitter、电影评分)中推断目标用户的私人信息(如位置、职业、爱好、政治观点)。此外，一系列研究工作已经证明，先进的人工神经网络可以作为一种对抗工具，从普通甚至模糊图像中检测

图像中的敏感信息，包括人们的年龄、关系和车牌号。图4(b)是推理攻击的一个说明。因此，加快针对机器学习辅助攻击的隐私保护方案的研究显得十分迫切。

3.2 基于机器学习攻击的隐私保护方案

在这一领域已经进行了一些初步的研究。为了保护隐私不受传统机器学习攻击，Liu等人 [21]设计了基于社区的信息共享方案，改变整体的时空特征，使基于聚类的隐私攻击不再起作用。

当涉及到深度学习时，问题变得更具挑战性。解决方案可能来自对深度学习本身更好的理解。一些研究人员最近发现，深度学习有局限性。具体来说，“它被证明容易受到一些良好设计的输入（对抗样本）的攻击” 。 Szegedy等人 [39]首次发现，在原始图像上叠加“不易察觉的噪声”会误导DNN进行错误的分类。然后，Goodfellow等人 [10]提出了“快速梯度符号法(FGSM)，可用于生成这类对抗样本”。可以在 [26, 17, 33]中找到产生此类噪声的其他算法。

根据 [39]，神经网络易受对抗样本攻击的主要原因是神经网络的线性性质。作者对DNN的对抗空间进行了形式化描述，这些对抗大多来自于机器学习技术本身。简单地说，机器学习被用作破坏机器学习分类器的工具。Kurakin等人 [17]专注于对抗训练以及如何将其扩展到大型数据集。Sharif等人 [34]提出了一种基于机器学习制造对抗样本的算法，使DNN检测系统无法在共享照片中寻找目标。另外，对抗样本的一个重要特点是其可移动性 [10]。这意味着，如果他们能够欺骗一个模型，他们通常很可能用一组不同的参数和架构来误导另一个模型 [39]。即使另一个模型是在不同的训练集或模型上训练的，这也是正确。这就引出了普遍扰动的概念。甚至有可能“生成能愚弄人类和计算机的对抗样本”。Elsayed等人 [23]利用机器学习构造对抗样本，从基于计算机视觉创建的模型转移到人类视觉系统。作者在没有利用模型架构的参数的情况下生成了对抗样本，然后使用机器学习模拟人类的视觉处理。

受到对抗样本思想的启发，研究者开始关注基于机器学习的对抗样本生成，以提高用户的隐私性，防止主要基于DNN的攻击。Liu等人 [20]提出了一种基于“Faster RCNN框架”的算法，该算法使用对抗样本来对抗自动检测。Jia等人 [17]提出了一个名为AttriGuard的两阶段框架来防御分类器发起的属性推理攻击。Liu等人 [20]研究了在机器学习系统中使用对抗例子的方案，这样它们就不能从图像中识别敏感信息。Oh等 [26]建立了博弈论框架，研究了对抗图像扰动对隐私保护的有效性。Li等人 [?]提出使用对抗扰动进行人脸去识别。Friedrich等人 [39]提出了一种保护隐私的医学文本可共享表示，用于反识别分类器。

3.3 基于机器学习攻击的隐私保护总结

以前，人们对隐私保护的普遍理解是防止人类对手知道某些人的敏感信息。例如，对图像中的人脸进行模糊处理是一个经过深入研究的课题。然而，最近情况发生了巨大变化。首先，数据量的增长已经达到了一个临界点，任何人都不可能用眼睛浏览所有内容。第二，因此，人们越来越依赖具有高级算法的机器来提取感兴趣的相关信息。第三，机器学习开源社区的蓬勃发展使得任何人都可以很容易地获得机器学习工具。这带来了一个新问题，即现在可以自动处理数据来推断敏感的用户信息，如个人身份、社会关系、和位置。事实上，最近恶意方利用机器学习作为发起新型隐私攻击的有效工具，尤其是针对社交媒体数据的攻击。因此，我们认为针对机器的隐私保护与针对人类的隐私保护一样重要。

基于机器学习的隐私攻击防御起来更具挑战性，主要有三个原因。首先，一般用户不知道最先进的机器学习方法在提取个人信息方面的能力。其次，在某些情况下，比如多媒体数据的隐私并不明显。第三，隐私威胁也产生于大规模收集和分析数据的组织和政府部门。因此, 我们需要防止机器学习算法自动挖掘隐私信息，不管是有意还是无意。

总之，针对快速发展的机器学习技术的隐私保护是我们在本文中讨论的三个类别中最具挑战性的任务。该方法利用了机器学习方法的弱点和局限性。虽然对抗机器学习已经初步解决了这一问题，但仍有许多研究问题需要进一步研究。.

4 未来的展望与发展

以前的重要工作集中于使机器学习算法具有不同的私密性，以保护训练集的私密性。然而应该意识到，机器学习作为一个整体，也为隐私研究(不仅仅是训练数据集)提供了强有力的工具，无论是从攻击还是防御的角度。

4.1 深度学习中的扰动

在深度学习中，扰动的目标是训练一个模型，同时确保DP涉及关于单个训练示例的信息。理论上，噪声可以添加到输入数据、模型参数(通过梯度更新)或模型输出中。在实际工作中，大多数工作提议将噪声注入梯度。这组方法的主要缺点是注入噪声的数量依赖于训练轮数，而且由于参数的数量很大，可能会累积太多的噪声。

直接向输入数据添加噪声是一种选择，但这类似于典型的大数据隐私问题，与深度学习没有密切关联。输出扰动和目标扰动似乎是未来的合理方向。

输出扰动是给机器学习系统的输出增加了噪声，例如预测阶段的对数。该方法具有快速、易于实现的特点。但是，由于对手的重复查询攻击，它可能会降级。因此，限制查询的数量很重要 [32]。一个可能的解决方案是在某些中间输出中使用输出扰动，例如PATE框架中的教师投票输出 [29].

目标扰动是差分隐私机器学习最有效的方法之一，该方法在目标函数中增加一个随机线性项。目标扰动在凸优化问题中得到了广泛的研究。最近，Iyengar等人 [13]提出了一种实用的差分私有凸优化算法，这是该技术走向实际部署的一大步。此外，Neel等人 [27]将该方法扩展到非凸优化问题。尽管传统机器学习方法取得了成功，但将目标扰动应用于深度神经网络仍存在一些困难:1)深度学习模型的目标函数大多是非凸的，没有封闭式表达式，敏感性计算困难;2)隐私保障隐式地基于损失的Hessian的秩1假设，难以验证;3)隐私保证只存在于优化问题的精确最小值(至少是 [13]中提出的近似最小值)，这在实际深度学习系统中是难以保证的。一种可能的解决方案是使用损失函数的凸逼近。然而，由于灵敏度较小，近似误差可能大于减小的扰动。预计在这个方向上会有更多有效的方法。

此外，除了干扰最终输出外，还可以在神经网络的中间层添加噪声。Lecuyer等人 [18]提出了在DNN中包含DP噪声层的PixelDP框架。虽然PixelDP的目的是“增强对抗样本的鲁棒性”，但这个想法可以进一步研究以服务于隐私保护。例如, PixelDP方案强制输出预测函数是DP提供的输入变化的少量像素(当输入是一个图像)。PixelDP的潜在扩展包括:1)对给定的不同输入样本执行DP，从而可以为训练集提供隐私保护，防止成员推理攻击;2)在自编码器的隐层中加入DP噪声。由于DP的后处理特性，自动编码器的输出仍然是DP。在 [18]中简要地提到了这一观点。但我们可以在不同的应用中进一步探索它。例如，我们可以通过使用这个自编码器与差分隐私保障来生成一个扰动版本来达到保护一个社交网络图像的目的。

4.2 基于机器学习的隐私攻击的保护方案：对抗样本

正如我们在第5节中讨论的，当机器学习被用作隐私攻击方法时，对抗样本是隐私保护的一种强有力的方法。尽管对这一主题进行了初步的工作，但仍有几个问题需要解决:

- 对抗样本生成方法分为两类攻击场景:白盒和黑盒。使用对抗样本来保护隐私的研究通常假设深度学习模型是已知的，使用白盒设置。在实践中，黑盒场景似乎是一个更现实的假设，例如，最新的黑盒对抗生成方法，如ZOO [6]， $\mathcal{N}$攻击 [19]和AdvFlow [7]，可以潜在地用于隐私保护。
- 在隐私和效用方面，仍然难以评估该机制的有效性。现有的作品使用机器学习输出(标签)的变化来评估隐私保护方法。我们需要提出更简明和更好的评估指标。
- 最近有一些研究将DP框架和对抗样本联系起来 [18]。PixelDP算法 [18]提出在网络架构的输入或任何中间层中添加一个DP噪声，以保证对对抗样本的鲁棒性。更详细地说，如果我们考虑“DNN的输入(例如图像)看成是DP中是数据库，个体特征(例如像素)看成是DP中的行”，随机输出预测函数

来加强DP可以保证预测对对抗样本的鲁棒性。由于输入变化被限制在“少量像素”，PixelDP不能有效地保护训练集的隐私 [18]。Phan等人 [18]提出了一种异构高斯机制(HGM)，该机制可以在训练数据中保留DP，同时对对抗样本提供可证明的鲁棒性。他们在 [31]中进一步提出了随机批处理机制，与HGM相比，该机制能够保留更高的模型效用，且对大型DNN和数据集更具可伸缩性。总的来说，DP、对抗样本和经过认证的鲁棒性之间的相互作用将是未来一个非常有趣的研究方向。

4.3 机器学习辅助的隐私保护: 生成对抗网络和变分自编码器

大量的非结构化数据(包括图像、视频、音频和文本)正在不断生成，政府和许多行业都在使用这些数据。根据国际数据公司的预测，到2025年，非结构化数据将占全球数据的80%左右。非结构化数据，尤其是图像和视频，往往包含着丰富的个人信息，在未来的隐私保护生态系统中扮演着关键的角色。而非结构化数据的私有数据发布问题将是未来研究的热点。我们希望GAN在这一领域发挥重要作用，因为它已经展示了在保护数据集中敏感信息的同时，为机器学习算法保持高实用性的能力。此外，GAN作为VAE的一部分，也可用于信号数据输入的隐私保护 (例如,一个图像)。在这种情况下，我们可以对原始数据输入进行编码，然后使用一些额外的隐私保护对其进行解码。

5 总结

本文调查了机器学习背景下关于隐私保护的文献。通过将现有的研究分为三组:(i)私有机器学习，(ii)机器学习辅助隐私保护(iii)防止机器学习攻击的隐私保护，我们全面回顾了这一主题的最新技术，并得出以下几点结论。

- 私有机器学习问题是近年来最受关注的问题。在这类研究工作中，许多人在分析过程中尝试使用差分隐私标准。但是，由于数据和隐私保护目标的复杂性，DP表示法不能提供全面的隐私评估。因此，如何定义新的隐私度量和表示法仍然是一个悬而未决的问题。
- 近来，关于机器学习帮助保护隐私的研究势头日益强劲。例如，使用GNN生成合成数据集为隐私保护研究开辟了新的方向，特别是对非结构化数据，如图像和视频。
- 针对基于机器学习的隐私攻击的保护方案的研究还处于起步阶段。但由于人工智能技术在未来网络的每个角落的扩散，预计它将在未来盛行。目前，这类领域的主流技术是对抗性样本和扰动技术。

我相信，本文将为与隐私和机器学习相关的研究问题提供有价值的启示。随着人们对这一主题的日益关注，我们期待在这一领域看到越来越多的研究成果。

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, and Li Zhang. On the protection of private information in machine learning systems: Two recent approches. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 1–6. IEEE, 2017.

[3] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

[4] Christopher M Biship. Pattern recognition and machine learning (information science and statistics), 2007.

[5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[7] Hadi M Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. *arXiv preprint arXiv:2007.07435*, 2020.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[11] Yulong Gu, Yuan Yao, Weidong Liu, and Jiaxing Song. We know where you are: Home location identification in location-based social networks. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2016.

[12] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618, 2017.

[13] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

[14] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

[15] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[16] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[18] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[19] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *International Conference on Machine Learning*, pages 3866–3876. PMLR, 2019.

[20] Bo Liu, Ming Ding, Tianqing Zhu, Yong Xiang, and Wanlei Zhou. Adversaries or allies? privacy and deep learning in big data era. *Concurrency and Computation: Practice and Experience*, 31(19):e5102, 2019.

[21] Bo Liu, Wanlei Zhou, Shui Yu, Kun Wang, Yu Wang, Yong Xiang, and Jin Li. Home location protection in mobile social networks: a community based method (short paper). In *International Conference on Information Security Practice and Experience*, pages 694–704. Springer, 2017.

[22] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, Tom H Luan, and Haibo Zhou. Silence is golden: Enhancing privacy of location-based services by content broadcasting and active caching in wireless vehicular networks. *IEEE transactions on vehicular technology*, 65(12):9942–9953, 2016.

[23] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor CM Leung. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, 6:12103–12117, 2018.

[24] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–21, 2014.

[25] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[27] Seth Neel, Aaron Roth, Giuseppe Vietri, and Steven Wu. Oracle efficient private non-convex optimization. In *International Conference on Machine Learning*, pages 7243–7252. PMLR, 2020.

[28] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3686–3695, 2017.

[29] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[30] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[31] Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning*, pages 7683–7694. PMLR, 2020.

[32] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. Sampling attacks: Amplification of membership inference attacks by repeated queries. *arXiv preprint arXiv:2009.00395*, 2020.

[33] Andras Rozsa, Ethan M Rudd, and Terrance E Boult. Adversarial diversity and hard positive generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–32, 2016.

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016.

[35] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[36] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[37] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017.

[38] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[40] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

[41] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security*, 12(5):1005–1016, 2016.

[42] Dayin Zhang, Xiaojun Chen, Dakui Wang, and Jinqiao Shi. A survey on collaborative deep learning and privacy-preserving. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 652–658. IEEE, 2018.