IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

2020 年 3 月 7 日接收，2020 年 4 月 5 日受理，2020 年 4 月 13 日出版，2020 年 5 月 4 日现行版本。

数字对象识别符 10.1109/访问。59609.88866868616

机器学习安全:威胁、对策和评估

薛明福 1，(IEEE 会员)，袁 1，吴 2，
张玉书 1 人(IEEE 会员)，刘伟强 3 人(IEEE 资深会员)
1 南京航空航天大学计算机科学与技术学院，南京 210016
2 南京优普赛网络安全技术研究院有限公司，中国南京 211100
3 南京航空航天大学电子与信息工程学院，南京 210016

通讯作者:薛明福和)

本研究得到了国家自然科学基金项目 61602241 和江苏省自然科学基金项目 BK20150758 的资助。

摘要近年来，机器学习因其技术上的突破而得到了广泛的应用。它在处理各种复杂问题方面取得了显著成功，并显示出接近人类甚至超越人类的能力。然而，最近的研究表明，机器学习模型容易受到各种攻击，这将危及模型本身和应用系统的安全。此外，由于深度学习模型无法解释的性质，这种攻击是隐形的。在本次调查中，我们系统地分析了机器学习的安全问题，重点关注对机器学习系统的现有攻击，相应的防御或安全学习技术，以及安全评估方法。本文不关注一个阶段或一种类型的攻击，而是涵盖了从训练阶段到测试阶段的机器学习安全的所有方面。首先，给出了对手环境下的机器学习模型，分析了机器学习受到攻击的原因。然后，将机器学习安全相关问题分为五类:训练集中毒；训练集中的后门；对抗性的示例攻击；模型盗窃；敏感训练数据的恢复。系统地分析了威胁模型、攻击方法和防御技术。为了证明这些威胁是现实世界中的真实问题，我们还回顾了现实世界中的攻击。最后，对机器学习系统的安全评估提出了几点建议。最后，提出了机器学习安全的未来发展方向。

索引术语人工智能安全、中毒攻击、后门攻击、对抗性示例、保护隐私的机器学习。

I. 介绍

近年来，机器学习技术取得了重大突破，并被广泛应用于图像分类、自动驾驶汽车、自然语言处理、语音识别和智能医疗等多个领域。在一些应用中，例如图像分类，机器学习的精度甚至超过了人类。机器学习也已经应用于一些安全检测场景，例如垃圾邮件过滤、恶意程序检测，这实现了新的安全特征和能力。

然而，最近的研究表明，机器学习模型本身面临许多安全威胁:1)训练数据

负责协调手稿审核并批准出版的副主编是张巍 。

中毒可能导致模型准确性降低或导致其他一般错误/特定错误攻击目的；2)训练数据中精心设计的后门可以触发系统的危险后果；3)测试输入中精心制作的干扰(对立的例子)会使模型出错；4)模型窃取攻击、模型反演攻击和隶属度推断攻击可以窃取模型参数或恢复敏感的训练数据。所有上述安全威胁都会给机器学习系统带来严重后果，特别是在安全和安全关键应用中，如自动驾驶、智能安全、智能医疗等。

近年来，机器学习安全引起了广泛关注[1]，[2]。关于深度学习算法的安全性有大量的研究工作

因为 Szegedy 等人[1]强调了深度学习算法中对抗性例子的威胁。然而，机器学习安全并不是一个新概念[3]，更早的工作可以追溯到 Dalvi 等人[4]在 2004 年。这些早期的工作，例如[4]，[5]，在垃圾邮件检测，PDF 恶意软件检测，入侵检测等背景下，研究了非深度机器学习算法上的所谓对抗性机器学习[3]。这些早期攻击中的大多数被称为逃避攻击，而少数其他攻击被称为中毒攻击。

基于这些问题，本文对机器学习的安全性进行了全面的综述。迄今为止，只有少数关于机器学习隐私和安全问题的评论和调查论文被发表。2010 年，Barreno 等人[6]回顾了早期对非深度学习算法的规避攻击，并在垃圾邮件过滤器上进行了说明。Akhtar 和 Mian [7]回顾了计算机视觉领域对深度学习的对抗性示例攻击。他们讨论对抗性示例攻击，并侧重于计算机视觉。袁等[8]对深度学习中的对立样例进行了综述，总结了对立样例的生成方法并讨论了对策。Riazi 和 Koushanfar [9]分析了可证明安全的隐私保护深度学习技术。他们讨论机器学习中的隐私保护技术，并关注基于密码原语的隐私保护方法。以上的综述著作都只关注一种类型的攻击，大多是对抗性的例子攻击。比格吉奥和花小蕾[3]对过去十年中对抗性机器学习的野生模式(也称为对抗性例子)进行了综述，包括早期非深度机器学习算法的安全性和最近计算机视觉和网络安全领域的深度学习算法。特别讨论了规避攻击和中毒攻击，并给出了相应的防御措施[3]。刘等[10]分析机器学习的安全威胁和防御。他们关注安全评估和数据安全。Papernot 等人[11]将机器学习中的安全和隐私问题系统化。特别是，他们根据三个经典的安全属性(即机密性、完整性和可用性)描述了攻击，同时他们从健壮性、责任性和隐私性方面讨论了防御措施[11]。

本调查与这些现有的少数综述/调查论文之间的差异总结如下:

1) 本文没有将重点放在一个阶段、一种攻击类型或一种特定的防御方法上，而是系统地涵盖了机器学习安全的所有方面。从训练阶段到测试阶段，所有类型的攻击和防御都以系统的方式进行了审查。
2) 给出了存在敌人时的机器学习模型，分析了机器学习受到攻击的原因。
3) 描述了威胁和攻击模型。此外，机器学习的安全问题被分为五类，涵盖了所有的安全威胁根据机器学习系统的生命周期，即训练阶段和测试阶段。具体来说，回顾和分析了五种类型的攻击:1)数据中毒；2)借壳；3)反例；4)模型窃取攻击；5)敏感训练数据的恢复，包括模型反演攻击和成员推理攻击。
4) 回顾和分析了基于机器学习系统生命周期的防御技术。此外，还分析了当前防御方法面临的挑战。
5) 对机器学习算法的安全性评估提出了一些建议，包括安全性设计、使用一组强攻击进行评估以及评估指标。
6) 提出了机器学习安全的未来研究方向，包括:真实物理条件下的攻击；保护隐私的机器学习技术:DNN 的知识产权保护；远程或轻量级机器学习安全技术；系统机器学习安全评估方法:这些对机器学习的攻击和防御背后的潜在原因。

本文的其余部分组织如下。存在对手时的机器学习模型，以及机器学习可能受到攻击的原因，将在第节中描述 II。第节回顾了威胁模型和攻击方法 III。第 3 节分析了防御技术和挑战 IV。机器学习算法的安全性评估将在第节中讨论 V。机器学习安全性的未来方向在第节中介绍 VI。我们在第一节结束这篇论文 VII。

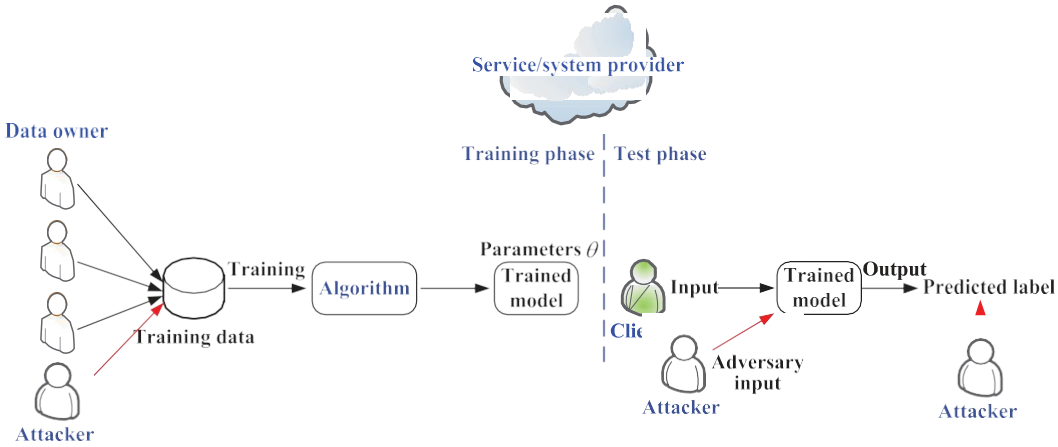
II. 对手面前的机器学习模型

A. 机器学习概述

一个机器学习系统的概述如图 2 所示。1。我们从以下几个方面来描述机器学习系统。

阶段:通常，机器学习系统可以分为两个阶段:1)训练阶段，其中算法从训练数据中学习以形成具有模型参数的模型；2)和测试阶段，在该阶段，将训练好的模型应用于特定任务，例如分类，以给出输入数据的预测标签。

算法:使用算法从训练集中学习，以获得带有参数的模型。我们将机器学习算法分为两类，神经网络算法和非神经网络算法。我们使用术语 NN 算法来代表深度神经网络(DNN)、卷积神经网络(CNN)、递归神经网络(RNN)和其他神经网络(NN)算法，这些算法在最近几年取得了重大突破，并显著提高了机器学习系统的性能。另一方面，我们使用术语非神经网络算法来表示其他



图一。概述机器学习系统，它说明了两个阶段，学习算法和不同的实体。

传统的机器学习算法，如支持向量机(SVM)、k-means、朴素贝叶斯等。

*对抗模型中的实体:一个正常的机器学习系统由以下实体组成，数据所有者、系统/服务提供商和客户端，而在对抗模型中，也有攻击者，如图。*1。数据所有者是大量训练数据的所有者，这些数据通常是私有的。系统/服务提供者是构建算法、训练模型，然后执行任务或提供服务的提供者。客户端是例如通过所提供的预测API来使用服务的用户。攻击者可以是外部对手，也可以是系统内部对其他实体的秘密信息感兴趣的好奇的人。

B. 为什么机器学习会被攻击

首先，机器学习的工作模式使其容易受到各种类型的攻击。在训练阶段，深度学习网络中训练过程的海量训练数据和计算复杂性导致:1)训练程序外包[12]；2)来自作为知识产权(IP)的第三方的预训练模型被集成到网络中；3)大量数据来自不可信用户或第三方，没有经过有效的数据验证。然而，上述工作模式也带来了新的安全威胁。

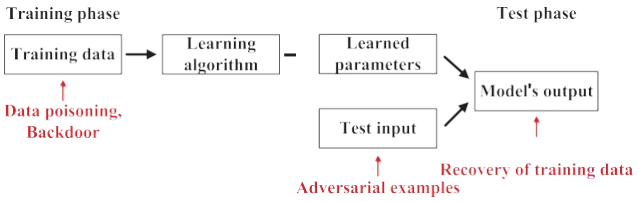
所谓的机器学习即服务也越来越多地被使用，其中机器学习模型在服务器或云上工作，而客户端可以通过预测API查询模型。用于训练该模型的大量数据通常是敏感的，并且该模型的参数具有很大的商业价值，因此是机密的。这些是测试阶段攻击者的目标。

第二，这些攻击背后的原因仍然是个悬而未决的问题。有一些关于这些对机器学习模型的成功攻击背后的原因的讨论，

然而，他们仍然缺乏共识。Goodfellow等人[2]指出，高维空间中DNN模型的线性是易受对抗性示例攻击的一个可能原因。研究[1]，[8]也认为训练数据的不完全性是对抗性例子存在的原因之一。他们得出的结论是，对抗性例子是测试集中概率较低的拐角情况，说明训练数据不够，不完整。Yeom等人[13]研究了过度拟合的效果以及对手恢复敏感训练数据或属性的影响。他们表明过拟合在使攻击者能够实施成员推理攻击中起着重要作用。由于机器学习模型无法解释的性质，这些攻击的本质原因，例如，对抗的例子是模型的缺陷还是固有属性，为什么敏感的训练数据可以通过正常查询恢复，仍然是公开的问题。

III. 对机器学习的攻击

在本节中，我们将回顾机器学习系统面临的威胁和攻击。如图所示。2迄今为止，沿着机器学习系统的生命周期的所有安全威胁可以分为五类:1)训练集中毒；2)训练集中的后门；3)对抗性例子攻击；4)模型盗窃；5)敏感训练数据的恢复(包括模型反演攻击和成员船推断攻击)。前两次攻击发生在训练阶段，后三次攻击发生在



图二。对机器学习系统的攻击。

测试阶段。我们将在接下来的部分中分别回顾这五种攻击，并在部分中讨论攻击 III-F。

A. 训练集中毒

旨在误导机器学习模型的预测的对训练集的恶意操纵被称为中毒攻击。研究表明，一小部分精心构建的中毒训练数据可以使机器学习模型的性能大幅下降。中毒攻击的概述如图 2 所示 3。在本文中，我们根据是否针对神经网络(NN)模型来划分中毒作品。表中给出了训练集中毒方法的总结 1 从研究方法的类型、研究对象、工作机制、研究效果、研究方法的优缺点等方面进行分析。

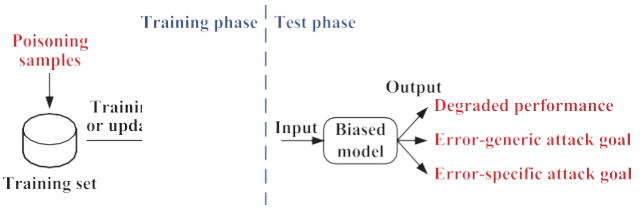


图 3。中毒攻击概述。

1) 针对非神经网络模型的中毒攻击针对异常检测或安全检测应用:机器学习已经广泛用于许多安全检测应用，例如异常检测和恶意软件检测。这些显然是投毒攻击的重要目标。Rubinstein 等人[14]对主干网络中基于主成分分析(PCA)-子空间方法的异常检测系统提出了三种中毒攻击。结果表明，通过引入一小部分中毒数据，检测器的性能急剧下降。这种方法简单而有效[14]，然而，它集中于二进制分类问题，这对于其他学习算法来说是不通用的。

李等人[15]使用边缘模式检测(EPD)算法在基于机器学习的入侵检测系统(IDS)上设计中毒攻击，称为慢性中毒攻击。该方法可以毒害几个学习算法，包括 SVM，LR 和 NB [15]。然而,[15]中的方法使用长期缓慢中毒程序，实施起来很复杂。

以生物特征识别系统为目标:机器学习技术也应用于自适应生物特征识别系统中，以适应用户生物特征的变化，例如老化效应。然而，更新过程可能被攻击者利用来危害系统的安全性[16]。Biggio 等人[16]提出了一种针对基于 PCA 的人脸识别系统的中毒攻击。通过提交一组精心设计的假面(即，中毒样本)并声称自己是受害者，系统模板会因自适应更新过程而逐渐被攻破。最后，攻击者可以用自己的脸冒充受害者。在[16]中，假设每个用户在系统中只存储一个模板，并且攻击者完全了解系统，例如特征提取算法、匹配算法、模板更新算法，甚至受害者的模板，这些在实际中很难获得。在[17]中，Biggio 等人针对更真实的人脸识别系统改进了上述攻击，其中系统为每个用户存储多个模板，使用不同的匹配算法，攻击者只有对受害者人脸图像的估计。然而，事实证明，攻击成功率取决于攻击者-受害者对[17]。

以 SVM 为目标:Biggio 等人[18]提出了针对支持向量机(SVM)的中毒攻击，其中注入了精心制作的训练数据，以增加 SVM 分类器的测试错误。他们使用基于 SVM 最优解的梯度上升策略来构建中毒数据。这种方法使用优化公式构建中毒数据，并且可以被内核化[18]，但是它需要算法和训练数据的完整知识。

目标聚类算法:聚类算法已广泛用于数据分析和安全应用，例如，市场细分、网页分类、恶意软件检测[19]。然而，群集过程本身可以被聪明的攻击者破坏。Biggio 等人[19]证明，攻击者可以通过在训练数据中引入少量中毒样本来破坏聚类过程。此外，这些中毒样本可以被有效地混淆和隐藏在输入数据中。在恶意软件样本聚类和手写数字聚类上对该方法进行了评估。Biggio 等人[20]提出了一种类似的中毒方法，该方法通过将带有中毒行为的特制中毒样本添加到训练数据[20]来针对行为恶意软件聚类。一般来说，这些方法[19]，[20]首先计算两个聚类之间的距离，然后在两个聚类之间插入中毒数据，这样会混淆两个聚类之间的界限。因此，聚类算法会错误地将两个不同的聚类合并成一个聚类[20]。这些方法[19]，[20]是通用的，可以攻击大多数聚类算法。然而，这些方法要求攻击者知道目标聚类算法、训练数据和特征空间等。

针对算法或处理方法:投毒攻击也可用于攻击特定的机器学习算法或处理方法。肖等人[21]研究了几种特征选择方法的中毒攻击，如岭回归、弹性网和最小绝对收缩选择算子(LASSO)。他们演示了 PDF 恶意软件检测，并表明中毒攻击可能会严重损害这些特征选择方法[21]。李等人[22]提出了对协同过滤系统的投毒攻击。它们表明，完全了解系统的攻击者可以构建

表 1。训练集中毒方法综述。

Targeting NN or non-NN	Target system/model	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Targeting non-NN models	Anomaly detection or security detection applications	Poisoning PCA-subspace based anomalies detection [14]; Chronic poisoning against IDS [15]	Poison the training data of the model	Reduce the performance of anomaly detection algorithms	Simple and effective [14]; Effective on several learning models [15]	Not generic [14]; Long-term slow poisoning and complicated to implement [15]
	Biometric recognition systems	Poisoning PCA-based face recognition systems [16]; Poisoning face templates [17]	Submitting a set of fake faces and claiming to be the victim	Replace the victim's template in the system	Stealthily and gradually	Require the knowledge of the target system and/or the victim
	SVM	Poisoning SVM [18]	Use a gradient ascent strategy based on the SVM's optimal solution	Increase the test errors of the SVM classifier	Use optimization formulation and can be kernelized	Need the full knowledge of the algorithm and the training data
	Clustering algorithms	Poisoning data clustering [19]; Poisoning behavioral malware clustering [20]	Poisoning attack based on the distance between clusters	Mislead clustering algorithms	Generic to clustering algorithms	Require the knowledge of the target algorithm
	Machine learning algorithms or processing methods	Poisoning feature selection [21]; Poisoning collaborative filtering [22]; Poisoning regression learning [23]	Optimization-based or statistical-based poisoning attacks	Mislead feature selection algorithms, collaborative filtering system or linear regression	Powerful and effective	Not generic, and/or require the knowledge of the target algorithm
	Online learning or data sanitization defenses	Poisoning breaks data sanitization defenses [24]	Optimization-based poisoning attacks	Bypass data sanitization defenses	Strong, effective under defenses	High computational overhead
		Online data poisoning [25] [26]	Optimization-based poisoning attacks	Reduce the performance of online learning	Can work under online learning scenarios	High computational overhead
Targeting NN models	\	Generative poisoning against NN [27]	GAN-based poisoning attack	Reduce the accuracy of the NN model	Generate poisoning data fast	Require interactions with the model frequently
		Poisoning of deep learning with back-gradient optimization [28]	Optimization-based poisoning attack	Increase multi-class classification error	Extend to multi-class problems and generalize well	High computational overhead
Specific application-oriented	\	Poisoning healthcare [29]; Poisoning graph-based recommender systems [30]; Poisoning crowd sensing systems [31]	Optimization-based poisoning attacks	Mislead healthcare system, recommender system or crowd sensing systems	Apply poisoning attacks to practical application scenarios	Require the knowledge of the target system

投毒数据以达到目的并模仿正常用户的行为以避免被注意到。在[21]、[22]中，他们将中毒攻击视为一个优化问题，并使用梯度上升策略来解决它。Jagielski 等人[23]调查了线性回归的中毒攻击，并使用数据的统计特性来生成中毒数据。这些方法[21]-[23]不是通用的，并且/或者要求攻击者具有目标算法的知识，例如特征集、学习算法等。

强数据中毒和在线学习投毒：中毒攻击和相应防御技术的演变也是一个螺旋式上升的过程。Koh 等人[24]提出了一种可以绕过现有数据净化防御的强中毒攻击。具体来说，他们将 poisoning 样本放置在彼此附近以逃避异常检测，并对优化问题设置一些约束以逃避检测。

现有的大多数中毒攻击工作都是在离线学习的情况下进行的。Wang 和 Chaudhuri [25]提出了针对在线学习情况的中毒攻击，其中训练数据以流的形式输入。他们将攻击公式化为优化问题，并使用梯度上升策略来解决它。此外，这种方法只修改输入流中特定位置的训练数据，以减少搜索空间，从而提高攻击效率[25]。类似地，Zhang 和 Zhu [26]提出了一种改进的在线中毒攻击，它需要有限的训练过程知识。他们将这种攻击公式化为随机最优控制，并提出了两种攻击算法，基于模型预测控制(MPC)攻击和基于强化学习的攻击。

强中毒攻击[24]在几种数据净化方法下都是有效的。[25]、[26]中的中毒攻击方法可以在在线学习场景中取得成功。然而，这些方法[24]-[26]通过解决优化问题来生成中毒数据，这需要昂贵的计算开销。

2) 针对神经网络模型的中毒攻击

近年来，神经网络的流行也导致了针对神经网络模型的中毒攻击。杨等[27]使用生成性对抗网络(GAN)

进行投毒攻击。具体地，自动编码器被用作生成器来生成中毒数据，而目标模型被用作鉴别器来计算中毒数据对模型的影响。该方法加快了中毒数据的生成，从而可以快速找到有效的中毒数据，但该方法需要频繁地与目标模型进行交互。

muoz-González 等人[28]提出通过使用反向梯度优化对深度学习进行中毒攻击。他们将中毒攻击从二进制算法扩展到多类问题。梯度是通过自动微分计算的。他们还逆转学习过程，以降低攻击的复杂性[28]。此外，研究表明，与测试阶段的对立例子类似，训练阶段的中毒例子也可以在不同的学习模型中很好地通用化[28]。然而，这种方法每次优化一个中毒数据，需要很高的计算开销[23]。

3) 特定应用场景下的投毒攻击

一些特定的有价值的应用场景成为了中毒攻击的目标。莫扎法里-克马尼等人[29]提出针对医疗数据集的中毒攻击。一般错误和特定错误都可能发生，这在医疗保健系统中是灾难性的。Fang 等人[30]提出了对推荐系统的中毒攻击。他们根据一个优化问题，用精心制作的评分生成虚假用户，并将其注入推荐系统。这样，攻击者可以将一个目标实例推荐给尽可能多的人。Miao 等人[31]提出了一个针对人群感知系统的中毒攻击框架。恶意工作者可以伪装成正常工作者来逃避检测，同时达到最大的攻击效用。在这些工作中，中毒攻击被应用于特定的场景，如医疗保健系统[29]，建议系统[30]和人群感应系统[31]。这些工作将中毒攻击形式化为不同的优化问题，以设计相应的中毒攻击策略。但是，这些方法[29] - [31]要求攻击者知道目标系统使用的学习算法或者关于训练数据的信息。

B. 训练集中的后门

最近，研究人员表明，攻击者可以创建一个隐藏在训练数据或预训练模型中的后门。

后门攻击的概述如图所示 4。后门不影响模型的正常功能，但是一旦特定的触发条件到来，后门实例就会被模型误分类为攻击指定的目标标签。这种后门攻击是隐形的，因为深度学习模型的无法解释的性质。后门攻击作品汇总如表所示 2。

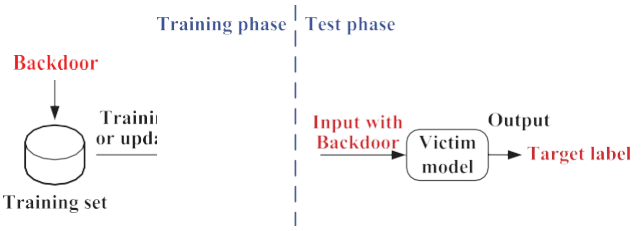


图 4。后门攻击概述。

1) 后门攻击

Gu 等人[12]提出了一个恶意训练的网络，命名为 BadNet。当特定的输入到达时，BadNet 会导致模型的不良行为。他们证明了 BadNet 在手写数字分类器和路标分类器上的有效性。纪等人[32]研究学习系统的后门。后门是由第三方提供的原始学习模块(PLM)引入的。一旦满足预定义的触发条件，集成到机器学习系统中的恶意 PLM 可以导致系统故障。他们在皮肤癌筛查系统上演示了这种攻击，而攻击者不需要关于该系统和培训过程的知识[32]。然而，在[32]中，攻击者直接操纵模型的参数来插入后门。这种假设在实践中很难满足。

Chen 等人[33]提出通过使用数据中毒对深度学习模型进行后门攻击。具体来说，中毒样本被注入到训练数据集中，以便植入后门。他们的攻击可以在弱攻击模型下工作，这意味着它不需要关于模型和训练集的知识[33]。只有 50 个中毒样本被注入，而攻击成功率在 90%以上[33]。廖等人[34]提出在 CNN 模型中通过注入隐形扰动进行后门攻击。特定的嵌入模式将被识别为攻击者定义的目标标签。

表二。后门攻击作品综述。

Scenario	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Backdoor attacks	BadNet [12]; PLMs [32]; Attack on deep learning [33]; Attack on CNNs [34] [35]; Attack on federated learning [36]	Add well-designed backdoor signals to the clean data; Manipulate the model's parameters directly	Target model misclassifies backdoor instances	Stealthy and don't affect the performance of the model	Need to participate in the training process, or re-train the model
Trojan attacks	PoTrojan [37]; Trojan attack on NN [38]	Add Trojan to the inner structure of NN [37]; Re-train the model to embed Trojan [38];	Target model produces malicious output	Without re-training [37]; Inverse the NN to generate the trigger [38];	Require the full knowledge of the model [37]; Requires re-training [38]

Barni 等人[35]提出了对 CNN 的后门攻击，其中他们在没有标签中毒的情况下破坏目标类的样本。他们评估了对 MNIST 数字分类器和交通标志分类器的攻击。

后门攻击甚至可以攻击最先进的安全训练模型和训练过程。Bagdasaryan 等人[36]阐述了对反馈式学习的后门攻击，反馈式学习被认为是一种安全的隐私保护学习框架。他们证明了恶意参与者可以通过模型替换将秘密后门功能引入全局模型。

在这些方法[12]、[33]-[36]中，攻击者首先将精心设计的后门信号加入干净的数据中，生成中毒数据，然后将中毒数据注入训练集，重新训练目标模型。重新训练后，在目标模型中嵌入特定的后门。这些方法[12]、[33]-[36]是隐蔽的，可以在不影响模型性能的情况下进行后门攻击，例如[34]中正常输入的精度下降小于 1%。但是，这些方法需要参与模型的训练过程，或者对模型进行重新训练。

2) 机器学习中的木马攻击

在一些文献中，这样的后门也被称为木马。刘等[39]提出了神经网络入侵防御系统中神经木马的初步概念，即在神经网络中插入恶意函数。他们提出了三种神经特洛伊缓解技术，即输入预处理、重新训练和异常检测(参见 IV-B 详情)，没有给出神经特洛伊实现的详细描述[39]。邹等人[37]在预训练的 NN 模型中提出神经级木马，命名为 PoTrojan。PoTrojan 保持不活动状态，只有在极少数情况下才会被触发，这可能会使模型产生恶意输出。具体来说，他们设计了两种触发器(单神经元 PoTrojans 和多神经元 PoTrojans)，并根据攻击者是否可以访问目标标签的训练数据来设计两种有效载荷[37]。Liu 等人[38]通过两步实现将木马插入 NN。首先，他们对神经网络求逆以产生特洛伊触发器。然后，他们重新训练神经网络模型，将恶意有效载荷注入神经网络。在这些方法[37]，[38]中，攻击者需要拥有目标神经网络的全部知识，这在实践中很难获得。

C. 对抗性示例攻击

对抗性示例是对攻击者精心构建的输入数据的干扰，以导致机器学习模型出错。Szegedy 等人[1]在 2014 年针对深度学习算法引入了术语“对抗性示例”。然而，类似的概念和方法要古老得多，它们被称为针对非深度机器学习算法的对抗性机器学习[3]，[4]。在这些早期的工作中，这些攻击被称为规避攻击，主要针对垃圾邮件过滤、恶意软件检测、入侵检测等。

对抗性示例攻击可以进一步分为两类[3]:错误-泛型攻击，它只是使模型出错；以及特定错误攻击，其目的在于使模型错误地将敌对样本识别为来自特定类别。表中给出了对抗性示例攻击的工作总结 3。

1) 对非深度学习算法的早期规避攻击

Dalvi 等人[4]首先强调了对手分类问题，其中对手试图使分类器产生错误的预测。此外，他们将分类表述为分类者和反对者之间的博弈，并提出了一个最优策略。Lowd 和 Meek [40]介绍了对抗性学习问题，其中对手试图通过发送大量查询来逆向设计分类器。这样，对手可以找到分类器不能识别的“恶意”实例。Nelson 等人[5]使用统计机器学习来攻击垃圾邮件过滤器，使其变得无用或实现重点目标。Barreno 等人[6]回顾了早期对机器学习系统的攻击，并提供了一种形式化的方法来描述攻击者和防御者之间的交互。他们在垃圾邮件过滤器 SpamBayes 上说明了他们的分类法。

Biggio 等人[41]将规避攻击转化为一个优化问题，并使用基于梯度下降的方法来解决它。他们通过增加对手的知识和能力来评估模型的安全性。他们证明了由基于梯度的方法生成的恶意示例可以逃避 PDF 恶意软件检测。rndi 和 Laskov [42]通过实验探索了逃避攻击下分类器(PDFRATE)的性能，发现即使在简单攻击下，PDFRATE 的分类性能也会显著下降。Demontis 等人[43]根据攻击者的不同知识和能力，在线性分类器 Android 恶意软件检测工具(Drebin)上实施了几种规避攻击。

在这些规避攻击中[4]、[5]、[40]-[43]，攻击者根据检测器的反馈修改对检测器影响最大的特征[56]。这些方法产生的恶意实例可以逃避安全相关应用程序的检测。但是，这些规避攻击需要攻击者知道目标系统的 tar- get 算法或特征提取算法所提取的特征[8]，这在实际中是很难获得的。另外，攻击成功率取决于攻击者掌握了多少知识。

与上述要求攻击者具有目标系统知识的方法相比，Xu 等人[44]和 Dang 等人[45]提出了不需要目标系统知识的规避攻击。根据恶意软件检测器返回的检测得分，Xu 等人[44]对恶意软件进行随机修改，找到一个可以规避检测但保留恶意行为的恶意实例。类似地，Dang 等人[45]提出了一种规避攻击，命名为 EvadeHC。EvadeHC 首先随机修改恶意软件以生成

表 3。对抗性示例攻击研究综述。

Scenario	Type/ Targets	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Earlier evasion attacks on non-NN models	Game theory	Adversarial classification [4]	Modify features that have the greatest impact on the detector	Malicious instances cannot be detected by the detector	Evade detection of security related applications	Require knowing the feature extraction algorithm
	Through reverse engineering the classifier	Adversarial learning [40]				
	Spam filtering	Nelson et al. [5]				
	PDF malware detection	Gradient-based evasion attack [41]; Šrndić and Laskov [42];				
	Android malware detection	Attack a linear classifier [43]				
	Malware detection (black-box)	Stochastic search-based attack [44]; EvadeHC [45]	Stochastic search-based; Heuristic search-based	Malicious instances cannot be detected by the detector	Generic; do not require knowledge about systems	Require iterations and feedback from the system
Adversarial example attacks on NN models	First work of adversarial examples on NN	Intriguing properties of NN [1]	Gradient-based or optimization-based adversarial example attacks	Adversarial examples mislead the model and cannot be perceived by humans	Perturbations are small and imperceptible; do not require knowing the feature extraction algorithm of the system	Attacks may fail in the real world
	One-step/one-shot methods	FGSM [2]				
	Take multiple small steps iteratively	BIM [46]				
	Perturbing only a few pixels	JSMA [47]; One pixel attack [48]				
	High-confidence adversarial examples	Carlini and Wagner [49], [50]				
	Universal adversarial perturbations	Universal perturbations [51]				
	Targeting deep learning based security detection	Malware classification [52]				
	In biometric authentication systems	Face recognition [53]				
In real world conditions	\	Road sign recognition attack [54]	Simulate physical conditions in adversarial examples generation process	Attack successfully in physical conditions	Robust to physical conditions	Perturbations are large and conspicuous
	\	Cellphone camera attack [46]				
	\	Face recognition system attack [53]				
	\	3D objects attack [55]				

恶意软件集合。然后，根据检测器返回的二进制结果(即拒绝或接受)，EvadeHC 使用爬山算法从恶意软件集合中搜索可以逃避检测的恶意软件。这些方法[44]，[45]对于二元分类问题是通用的，但是这些方法需要多次迭代才能获得有效的对立例子。因此，这些方法的计算开销很高。此外，这些攻击需要检测器的反馈，这限制了实用性。

2) 对深度学习的对抗性示例攻击对深度学习算法的对抗性示例攻击的流水线如图所示。5。Szegedy 等人[1]首先发现 dnn 对图像微小的难以察觉的扰动的对立例子是惊人地敏感的。这种对立的例子攻击会使 DNN 模式被错误分类。

文献中较早的对抗性攻击方法称为一步/一击方法，例如 Goodfellow 等人提出的快速梯度符号方法(FGSM)[2]。在后面的改进工作中，研究人员试图迭代地采取多个小步骤，以调整每步后的方向，而不是采取单个大步骤来增加分类器的损失[7]，例如，基本迭代法(BIM)[46]。

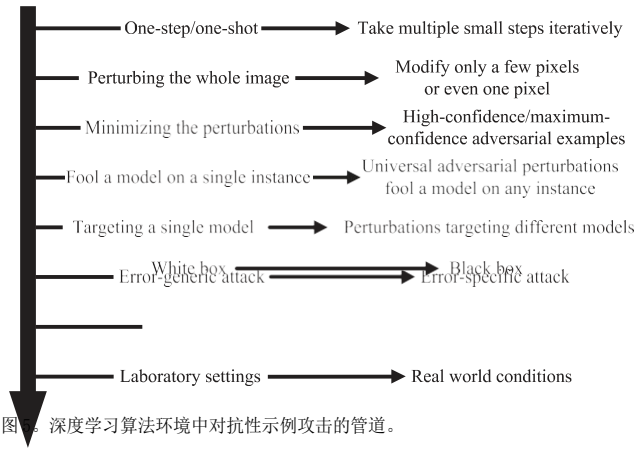


图 5。深度学习算法环境中对抗性示例攻击的管道。

另一个对立的例子试图只干扰图像中的几个像素，而不是整个图像[7]，例如，基于雅可比的显著图攻击(JSMA) [47]。Papernot 等人[47]提出了一种对抗示例生成方法，其中对手只需要知道模型的结构。作为一个极端的例子，苏等人[48]

提出一种方法，通过修改图像的一个像素来欺骗分类器。

请注意，虽然最初的敌对例子是最小的扰动，但在某些情况下，更合理的假设是对手希望最大化分类器对错误预测的信心，而不仅仅是最小化扰动[3]。原因是，最初对对抗性例子的工作旨在分析深度学习算法对最小扰动的敏感性，然而，为了分析深度学习算法在攻击下的安全性，使用最大置信度对抗性攻击更合理，它可以反映算法在更强大攻击下的安全性[3]。例如，Carlini 和 Wagner [49]，[50]表明，最近针对最小扰动攻击的几种防御技术可以被高置信度的对抗示例绕过。因此，他们建议在安全评估中使用高可信度的对抗性例子[49]，[50]。

在引入扰动以每次在单个图像上欺骗网络的方法(例如，FGSM [2]，Deep- Fool [57])之后，Moosavi-Dezfooli 等人[51]提出了可以在任何图像上欺骗网络的通用对抗性扰动。

尽管早期的工作是针对单个模型计算扰动，但是更强大的攻击可以产生跨不同模型的扰动[7]。研究表明，不同模型之间，尤其是具有相似结构的模型之间存在可移植性。

另外三个研究方向是从一般错误攻击到特定错误攻击，从白盒场景到黑盒场景，从实验室条件到真实世界条件。例如，Papernot 等人[58]提出了一种黑盒对抗性攻击策略，攻击者根据选择的输入观察 DNN 的输出，然后建立目标 DNN 模型的替代模型。他们使用替代模型来产生对立的例子，这些例子在目标模型中也被发现是有效的[58]。

作为特例，我们讨论两个具体的应用。一方面，深度学习已经在一些安全应用中得到应用。然而，对立的例子已经被用来损害这些基于深度学习的安全应用。例如，Grosse 等人[52]提出了针对基于 DNN 的恶意软件分类的对抗性示例攻击。另一方面，深度学习也已成功应用于生物认证系统，如人脸识别、语音控制系统等。然而，通过生成精心制作的对抗性示例，攻击者可以使模型错误地将攻击者识别为合法用户，从而获得合法用户的权限(特定错误攻击)，或者可以使攻击者规避系统识别(一般错误攻击)[53]。

在这些著作[1]、[2]、[46]-[53]中，对抗性例子主要是基于梯度或最优化生成的。基于梯度的方法计算梯度损失函数，并根据梯度将扰动添加到输入数据，以生成对立的例子。基于优化的方法将对立的实例生成问题转化为优化问题。优化的目标是对立实例中的扰动不仅会误导模型，而且不会被人类察觉。与早期对非深度机器学习算法的攻击相比，这些对深度学习算法的对抗性示例攻击不需要攻击者知道目标系统使用的特征提取算法。此外，生成的对抗样本中的扰动较小，不易察觉，可以获得较高的攻击成功率。然而，由于许多物理因素的影响(例如，角度、距离等。)，这些数字对抗的例子在现实世界条件下都失败了或者攻击成功率很低[59]。

也有几个有用的生成对立范例的工具箱，如 Cleverhans [60]，AdvBox [61]，Adversarial-Playground [62]，可以促进这方面的研究。由 Papernot 等人[60]开发的软件库 Cleverhans 提供了不同的对立示例生成方法的标准实现。AdvBox 是百度开发的一个工具包，支持 TensorFlow、PaddlePaddle 和 Caffe2 框架，用于生成深度学习模型的对抗示例[61]。Adversarial-Playground 是由 Norton 和 Qi [62]开发的基于网络的可视化工具。它可以实现针对 CNN 的常见对抗性示例生成方法。

3) 真实世界条件下的对抗性攻击上述对抗性攻击在实验室条件下进行了演示。为了让社区相信对抗性的例子在实践中是一个真正的问题，一些研究人员已经在实际的真实世界条件下说明了他们的对抗性攻击[7]。我们将从以下四个方面对这些作品进行回顾和分析:1)手机摄像头攻击；2)人脸识别系统攻击；3)路标识别攻击；4)和 3D 物体攻击。

路标识别攻击:物体识别是自主车辆的一项重要任务，需要识别路标、行人等。然而，Evtimov 等人[54]证明，在各种物理条件下，如视角、距离和分辨率的变化，对立的例子是稳健的。他们提出了两种实用的攻击方法[54]。第一种是海报打印攻击，攻击者将 C&W 攻击和其他算法生成的对抗性示例打印为海报，然后覆盖在真实的路标上。第二种是贴纸扰动攻击，攻击者将扰动印在纸上，然后贴在真实的路标上。这些物理条件的扰动可以成功地使模型分类错误，例如，停止标志被认为是限速标志[54]。

手机摄像头攻击:Kurakin 等人[46]打印了 FGSM、BIM 等生成的对抗性示例。

然后，他们用手机拍下印刷的对抗性例子。最后，他们使用 TensorFlow Android 相机演示对这些图像进行分类。结果显示，这些图像中的大多数都被错误分类[46]。这表明对抗例子在打印和拍照下是鲁棒的。

人脸识别系统攻击:人脸识别是计算机视觉中的一项重要技术，已经广泛应用于视频监控和门禁系统中。Sharif 等人[53]提出了在物理条件下工作的对抗示例，这使得攻击者能够逃避系统识别或冒充他人。攻击者打印并佩戴具有附加扰动的眼镜框。然后通过人脸识别系统验证攻击者。事实证明，佩戴这种眼镜的攻击者会被错误地识别为另一个人，从而避免人脸识别系统的检测[53]。

3D 对象攻击:现实世界中的 3D 对象是一个很难生成对抗实例的目标，因为它涉及许多角度。Athalye 等人[55]提出了一个对转换的期望(EOT)框架，该框架可以在不同的对象转换上构造对立的例子，从而可以打印对立的 3D 对象。他们的实验表明，一只 3D 打印的海龟会被 ImageNet 归类为步枪。

这些对抗性的示例攻击[46]、[53]-[55]考虑了各种物理因素的影响，如视角、距离、光照等。，使生成的对立实例对物理条件具有鲁棒性。因此，生成的对立范例可以在真实的物理条件下成功。然而，与数字对抗范例相比，这些物理对抗范例中添加的干扰更大、更明显，容易被视觉注意到。

D. 模型提取攻击

最近的研究表明，对手可以通过观察输出标签和关于所选输入的置信度来窃取机器学习模型。这种攻击也称为模型提取攻击或模型窃取攻击，已经成为一种新兴的威胁。模型提取攻击工作的总结如表所示 4。

Tramèr 等人[63]首先提出了模型提取攻击，即攻击者试图通过多次用户查询来窃取机器学习模型。当通过预测 API 输入正常查询时，模型将返回一个具有置信度的预测标签。基于这一服务，他们展示了对三种模型的模型提取攻击:逻辑回归、决策树和神经网络[63]。两个在线机器学习服务用于评估，Amazon 和 BigML。

Yi 等人[64]提出了一种基于深度学习构建功能等效模型的模型窃取方法。它在黑盒场景中工作，在黑盒场景中，广告只能从目标模型中获得预测的标签，并使用深度学习进行推断，然后建立一个等效模型[64]。具体来说，他们使用输入数据来查询目标模型，并使用目标模型返回的结果来标记输入数据[64]。标记的数据用于训练与目标模型具有相似功能的模型。Chandrasekaran 等人[65]表明，模型提取类似于主动学习。他们将模型抽取公式化为查询合成主动学习，提出了无辅助信息的模型抽取攻击。这些方法[63]-[65]通过对目标模型进行黑盒访问来训练与目标模型相似的模型，这不需要攻击者了解目标模型。然而，现有的模型提取攻击需要多次查询目标模型。如果系统限制查询的数量，这些模型提取攻击可能无法完成。

王和龚[66]的目的是通过使用一个学习器来窃取机器学习模型的超参数。该方法将模型的梯度设置为 0，然后通过求解线性方程计算模型的超参数[66]。方法[66]证明了模型提取攻击可以从许多机器学习模型中窃取超参数，例如 SVM、逻辑回归、岭回归和神经网络。然而，这种方法要求攻击者知道学习算法、训练数据等。Milli 等人[67]提出了一种算法，通过查询特定输入的目标模型的梯度信息来学习模型。结果表明，梯度信息可以快速揭示模型参数。他们得出结论，梯度是一个比预测标签更有效的学习原语[67]。然而，这种启发式方法引入了很高的计算开销，并且他们仅在两层神经网络上评估他们的模型提取攻击。

E. 敏感培训数据的恢复

除了上述模型提取攻击之外，对机器学习的其他两种隐私相关的攻击是:(I)成员推理攻击，其中攻击者试图确定在训练模型时是否使用了特定的样本数据；(二)模型反演攻击，攻击者推断出训练数据的一些信息。与模型提取攻击类似，隶属度推理攻击和模型反转攻击也是针对流行的机器学习即服务。这三种对机器学习模型的隐私相关攻击如图。6。表中总结了恢复敏感训练数据的攻击 5。

1) 模型反转攻击

弗雷德里克松等人[68]首先提出了遗传药理学任务中的模型倒置攻击。通过使用黑盒访问和关于患者的辅助信息，攻击者可以恢复患者的基因组信息。弗雷德里克松等人[69]进一步提出了利用预测的置信度值的模型反演攻击。他们演示了对基于决策树的生活方式调查和基于神经网络的面部识别的攻击。例如，在面部识别模型中，

表 4。模型提取攻击作品综述。

Approaches	Working mechanism	Effect	Advantages	Disadvantages
Tramèr et al. [63]; Shi et al. [64]; Chandrasekaran et al. [65]	Extraction through prediction APIs; Train a model similar to the target model; Use active learning [65]	Can generate a model with the same functionality as the target model	Extract model parameters without the knowledge of the target model	Need to query the target model frequently
Wang and Gong [66]	Calculate hyperparameters by solving linear equations	Steal the hyperparameters of models	Suitable for various machine learning algorithms	Require the knowledge of the target algorithm and the training set
Milli et al. [67]	Query gradients of a model to chosen inputs, and heuristic model reconstruction	Can reconstruct the target model	More efficiently than querying labels based methods	High computational overhead and only evaluate on a two-layer NN

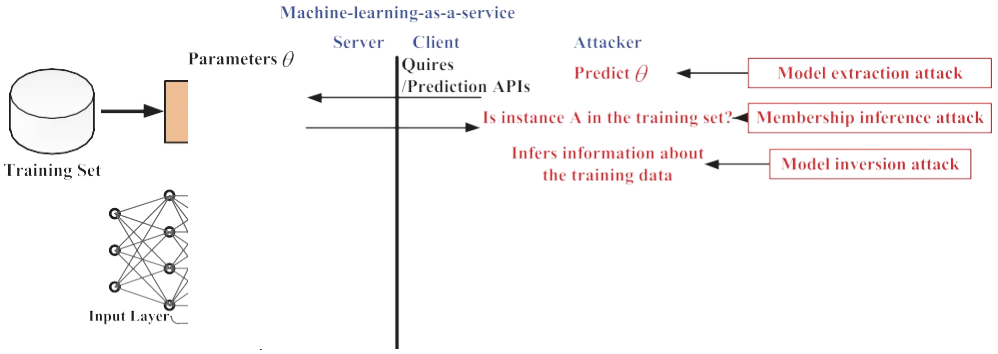


图 6。对机器学习模型的三种隐私相关攻击的概述:模型提取攻击、成员推理攻击和模型反转攻击。

表 5。针对恢复敏感训练数据的攻击摘要。

Approaches	Working mechanism	Effect	Advantages	Disadvantages
Model inversion attack [68] [69]	Estimate sensitive attributes through black-box access to the model	Infer sensitive attributes of the training data	Can infer sensitive attributes of data in sensitive applications	Require the knowledge of non-sensitive attributes of the data
Membership inference attack [70]	Train a model to distinguish training data and non-training data	Infer whether a given data is in the training data or not	Generic	Depends on the overfitting of the model
GAN based strong attacks [71]	GAN-based privacy information stealing	Can steal private information from collaborative deep learning	Generic; make differential privacy ineffective	The attacker needs to be an insider of the collaborative deep learning framework

知道用户姓名的攻击者可以恢复用户面部的可识别图像[69]。[69]中的方法比[68]中的方法具有更少的误报，然而，这些方法[68]、[69]需要攻击者知道数据的非敏感属性，这在实践中可能难以获得[72]。

2) 成员推理攻击

刘等人[73]阐述了认知系统中的安全威胁。具体来说，它们表明攻击者可以通过只使用模型的公共可访问服务来访问机密的训练数据或复制处理模型。Shokri 等人[70]提出了所谓的成员推理攻击，其中对手可以估计给定的数据是否在目标模型的训练集中。特别地，他们使用目标模型对训练和非训练数据的预测来训练成员推理模型[70]。根据目标模型的输出，生成的隶属度推理模型可以识别目标模型对其训练数据和未用于训练的数据的预测。[70]中提出的成员推理攻击是通用的，但成员推理攻击的成功取决于模型的过拟合[13]，[70]。如果是一个很好的泛化模型，隶属度推理攻击的成功率很低。

3) 基于 GAN 的强攻击

为了保护用户的隐私数据，研究人员最近提出了协作深度学习框架，其中每一方都在本地训练他的模型，只有一小部分参数子集被共享。差分隐私也被引入来混淆保护的参数[74]。然而，Hitaj 等人[71]提出了一种基于生成敌对网络(GAN)的强攻击，它可以打破上述分布式或联合框架。对手训练 GAN 以生成私有训练集的等价样本，其中生成的样本具有与私有训练集相同的分布。结果表明，微分

基于隐私的协作深度学习框架在面对这种攻击时是无效的[71]。然而，[71]中的攻击者需要是协作式深度学习框架的内部人员，该框架具有从服务提供商获得模型参数的权限。

F. 关于攻击的讨论

在最近十年，对机器学习的攻击大多是对抗性示例攻击，而其他四种类型的攻击明显较少。其中，对图像的例证研究居多，而对言语和文本的例证研究相对较少。近年来，隐私相关攻击不断出现，并受到越来越多的关注。我们将机器学习攻击的趋势总结如下：

1) 攻击趋向于更实际、更真实的物理条件，例如在第节中描述的在真实世界条件下的对抗性示例攻击 III-C.3。例如，攻击手机或监控摄像头中的人脸识别系统，或者攻击无人驾驶汽车的路标识别系统。
2) 攻击越来越强，甚至可以颠覆人类的常规认知。例如，最先进的对抗示例不仅可以使模型输出错误的预测(例如，错误地将停止标志识别为速度限制标志)，还可以使模型不知道这是路标[75]或者不知道这是人[76]。例如，通过在衣服上粘贴印刷的敌对示例图片，人类可以在人检测器前隐藏自己[76]。这种类型的攻击可以用来躲避监控系统。
3) 针对生物认证系统的攻击层出不穷。在智能物联网时代，认证和控制是两个关键特征。有许多基于生物特征的认证和控制系统，例如基于指纹和基于语音的系统。然而，上述攻击可以成功地突破这些生物认证系统，从而威胁到控制系统的安全。例如，智能语音伪造可以欺骗自动说话人确认系统，从而侵入系统。

IV. 辩护

在本节中，我们将根据机器学习系统的生命周期来回顾和分析针对上述攻击的防御措施。现有的机器学习防御技术，涵盖了针对上述五种安全威胁的对策，总结于图。7。针对上述五种攻击的防御方法在第一节中进行了分析 IV-A 部分 IV-E，分别是。我们在第 5 节总结了防御技术 IV-F。

A. 防御中毒攻击

我们根据中毒攻击是否针对神经网络模型来讨论对中毒攻击的防御。针对中毒攻击的现有防御技术的总结在表中给出 6。

1) 非神经网络模型中对中毒攻击的防御

异常检测或与安全相关的检测中的防御：Rubinstein 等人[14]提出了一种防御技术，称为解毒剂，可以防止对异常检测器的中毒攻击。解毒剂使用稳健的统计数据来减轻离群值的影响，并可以拒绝中毒的样本。Biggio 等人[77]将中毒攻击缓解视为异常值检测问题，这种问题数量很少，并且与正常训练数据相比，分布发生了变化。因此，他们使用 Bagging 分类器，这是一种集成方法，可以减轻训练集中这些异常值(中毒样本)的影响。具体来说，他们使用不同的训练数据来训练多个分类器，并结合多个分类器的预测来减少训练集中离群值的影响[77]。他们评估了垃圾邮件过滤器和基于网络的入侵检测系统的集成方法，以防止中毒攻击[77]。然而，训练多个分类器将引入显著的开销。Chen 等人[78]针对恶意软件检测系统中的中毒攻击提出了一种防御技术，称为 KUAFUDET。KUAFUDET 使用一个自适应学习框架，并使用一个检测器来过滤可疑的假阴性，然后将其输入到训练程序中[78]。这些方法[14]，[77]，[78]可以提高学习算法的鲁棒性，减轻离群点对训练模型的影响，但这些方法主要关注二分类问题，引入额外开销。

为辩护：张和朱[79]提出了一个基于博弈论的分布式辩护。他们使用博弈论来分析攻击者和学习者之间的利益冲突。纳什均衡用于预测敌对环境中学习者的结果[79]。这种方法可以防止错误的更新，并防止中毒数据降低分布式 SVM 的性能。然而，这种基于博弈论的防御技术需要昂贵的计算开销。

算法和处理方法中的防御：刘等人[80]提出了一种针对中毒攻击的稳健线性回归方法。该技术首先使用低秩矩阵分解，然后使用主成分回归来修剪中毒样本。Jagielski 等人[23]提出了一种用于回归学习的防御算法，名为 TRIM。TRIM 使用迭代方法来估计回归参数，而 TRIM 损失函数用于移除和隔离可疑中毒点。Steinhardt 等人[81]通过使用异常值移除和风险最小化来构建针对中毒攻击的防御方法。Baracaldo 等人[82]提出了一种基于数据起源的防御技术，用于

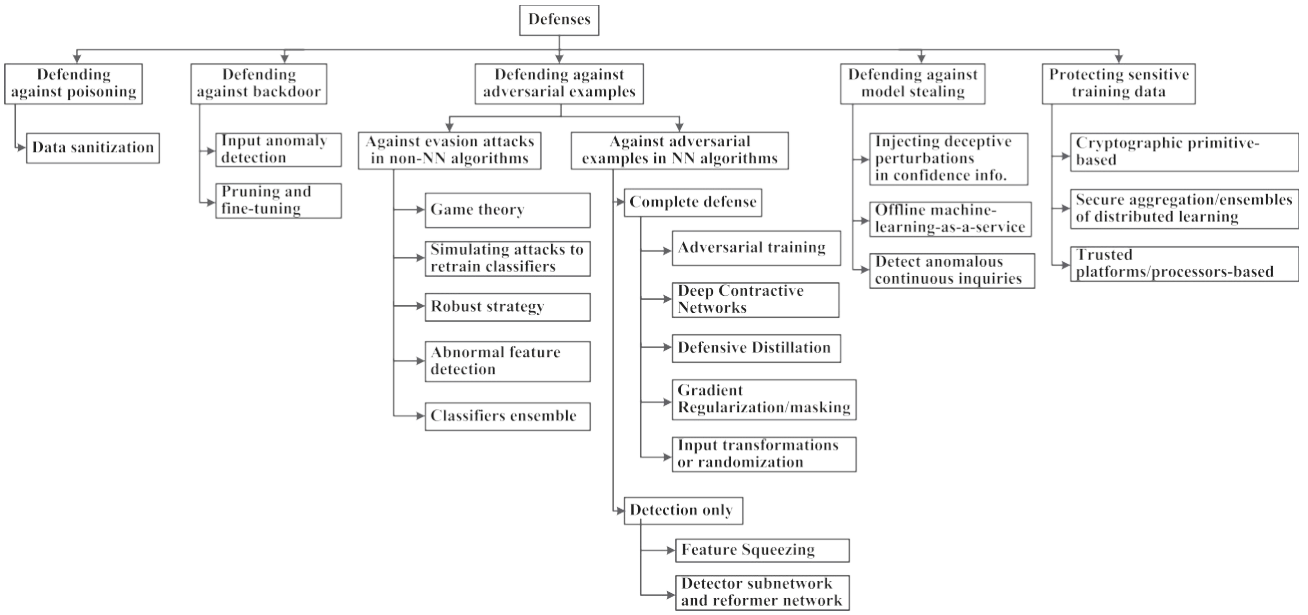


图 7。机器学习防御技术综述。

表 6。中毒攻击防御技术综述。

Targeting NN or non-NN	Target	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Defenses in non-NN models	Anomaly detection or security detection	Defending anomaly detectors [14]; In adversarial classification tasks [77]; In malware detection systems [78]	Using robust statistics; Bagging ensembles; Using self-adaptive learning and introducing a camouflage detector	Mitigate the influence of outliers	Improve the robustness of learning algorithms	Mainly focus on binary classification problems; Additional overhead
	SVM	Game-theoretic defense in distributed SVM [79]	Game theory and rejection method	Reject malicious updates	Prevent poisoning and wrong updates	Not generic; High computational overhead
	Algorithms and processing methods	Robust linear regression [80]; Regression learning [23]; Certified defenses [81]; Data provenance [82]	Low-rank matrix factorization and principle component regression; Iterative method; Outlier removal and risk minimization; Data provenance based on contextual information	Remove poisoning data	Improve the robustness of learning algorithms	Not generic
Defenses in NN models	\	In NN [27];	Check the loss of the model	Detect the poisoning attack	Simple, generic	Not fully evaluated in the experiment
		In collaborative deep learning systems [83]	Statistical-based: identify features with abnormal distributions	Automatically identify malicious users	Defend against poisoning attacks without affecting model performance	Defense performance is affected by the number of malicious users
Defenses in special applications	\	In healthcare [29]	Monitor the accuracy deviations on the training set	Detect poisoning data	Generic	High computational overhead

在线和重新培训的应用程序。实际上，他们利用训练数据的来源和转换信息来检测中毒点。这些方法[23]、[80]-[82]可以提高学习算法的鲁棒性，但这些方法不是通用的，只适用于特定的算法。

2) 神经网络模型对中毒攻击的防御

杨等人[27]在计算模型损失的基础上，提出了一种针对神经网络中毒攻击的对策。

引入的损失大于阈值的输入数据被认为是可疑数据。这种方法[27]简单且通用，但是它们仅呈现一个简单的检测结果，而没有完全评估防御方法。Shen 等人[83]提出了一个名为 AUROR 的系统来保护协作式深度学习系统。由于 poisoning 数据对模型学习到的特征分布有很大影响，AUROR 通过识别异常特征来过滤掉可疑用户[83]。AUROR 防御中毒攻击时不会影响目标模型的性能，但会影响防御性能

表 7。后门攻击防御概述。

Approaches	Working mechanism	Effect	Advantages	Disadvantages
Activation clustering [84]; Fine-Pruning [85]; Neural cleanse [86]; Strengthen the NN models [39]	Activation clustering-based detection; Fine-pruning; Neuron pruning, input filter, and unlearning based model patching; Adding an input preprocessor, input anomaly detection, and continuing re-training.	Detect and mitigate the backdoors	Generic for most DNNs	High computational overhead
STRIP [87]	Perturb the input and observe the prediction	Detect the backdoors	Run-time detection	May fail when facing adaptive attacks

这种方法的效率受恶意用户数量的影响。

3) 特殊应用场景下的中毒攻击防御

莫扎法里-克马尼等人[29]提出了一种对策，通过监控训练集的精度偏差和添加数据的数量来保护医疗保健系统。此方法是通用的，可以针对不同的目标模型提供防范中毒攻击的保护。然而，这种方法需要定期训练模型[29]，这导致了很高的计算开销。

B. 防御后门攻击

桌子 7 总结了针对后门攻击的防御措施。Chen 等人[84]提出了一种通过检测和移除后门来保护的激活聚类方法。所提出的激活聚类方法可以检测出有倾向性的训练数据，即使在多种后门情况下也是如此。Liu 等人[85]研究了对抗后门攻击的两种防御方法，称为修剪和微调。然后，他们提出了一种结合修剪和微调的方法，称为精细修剪，以减轻后门的影响。Wang 等人[86]通过三种技术来识别和缓解后门，即使用输入过滤器来识别带有触发器的输入，使用神经元修剪来修补模型，以及使用基于无学习的模型修补。Liu 等人[39]通过三种方法加强了针对特洛伊木马的 NN 模型:使用和决策树的输入异常检测；添加输入预处理器，即根据合法训练数据训练的自动编码器，从而可以基于输入数据的分布对输入数据进行预处理；和继续再训练，这可以使模型“获得”特洛伊。这些防御方法[39]、[84]-[86]适用于大多数 dnn，但这些方法在检测和缓解后门时需要昂贵的计算开销。

高等人[87]提出了一种用于的实时检测技术，命名为 STRIP。这个想法是扰动输入，观察熵方面的预测随机性。如果是特洛伊输入，这种扰动的预测几乎不变[87]。如果是干净的输入，这种扰动引起的预测变化会很大。这种方法速度很快，可以在运行时检测到木马，但这种方法在面对适应性攻击时可能会失败[88]。

C. 对抗敌对例子攻击的防御

我们讨论了基于神经网络模型的对抗实例攻击的防御方法。表中列出了对抗实例的抗辩事由 8。

1) 非深度学习算法背景下的早期规避攻击防御

早期针对规避攻击的防御工作主要是针对垃圾邮件过滤、恶意软件检测和 IDS 环境中的非深度学习算法。Dalvi 等人[4]使用博弈论来描述对手的分类问题，它可以根据对手的策略产生最优的分类器。Nelson 等人[5]在垃圾邮件过滤的背景下提出了两种防御技术，测量每封电子邮件在有和没有该电子邮件的情况下对系统性能的影响，并使用动态阈值设置，以便排名不会因攻击导致的分数变化而改变。

在对立的机器学习场景中，通常期望分类器是稳健的，对数据分布的变化不敏感[89]。Glober- son 和 Roweis [90]介绍了一种稳健的学习策略，通过利用博弈论分析稳健性来避免特征过度加权。特别是，他们开发了对特征删除有弹性的分类器，并在手写数字识别和垃圾邮件检测任务中进行评估。类似地，kocz 和 Teo [89]提出了两种方法，平均分类器和特征重新加权以提高鲁棒性，并对垃圾邮件分类任务进行了评估。Biggio 等人[91]通过实验评估了随机子空间方法和 bagging 这两种技术在对抗性机器学习下提高线性分类器鲁棒性的有效性。rndi 和 Laskov [92]提出了基于分层文档结构的恶意 PDF 检测，并评估了该方法在几种规避攻击下的鲁棒性。Demontis 等人[43]提出了一种用于 Android 恶意软件检测的对手感知安全学习范式，可以减轻规避攻击的负面影响。

在[93]中，李等人提出了一种迭代重训练方法来提高分类器的鲁棒性。他们通过反复向训练数据中添加可以逃避检测的恶意实例来重新训练分类器。迭代重新训练方法最大限度地降低了规避攻击的风险，并提高了分类器抵抗规避攻击的能力[93]。

表 8。对抗实例的抗辩摘要。

Targeting NN or non-NN	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Defenses in non-NN models	Game theory [4]; Defenses for Spam filter [5]; Feature weighting [89]; Feature deletion [90]; Multiple classifier system [91]; Malicious PDF detection [92]; Android malware detection [43]; Binary Domains [93]; SecDefender [94]; Feature Selection [95]; SecureDroid [96]; Conserved features [97]; Data transformations [98]	Game theory; Measuring the impact of each email and using dynamic threshold; Feature weighting; Feature deletion; Using multiple classifiers; Structural properties difference detection; Adversary-aware approach; Iterative re-training; Classifier re-training and security regularization; Feature selection; Using conserved features; Data transformations	Improves the ability of the model to resist evasion attacks	Simple and effective	Not generic; Mainly focus on binary classification problems
Defenses in NN models	Adversarial training [2], [99]	Adversarial training	Improve the robustness of the target model	Simple, effective, and generic	Increase training overhead
	Saddle point formulation [100]; Deep Contractive Network [101]; Defensive distillation [102]; Input gradient regularization [103]	Using saddle point formulation; Using a smoothness penalty; NN distillation; Input gradient regularization	Mitigate the effects of adversarial examples on the model	Generic to most NNs	May increase the training complexity or may be defeated by stronger attacks
	Input transformations [104]; Randomization [105]	Apply image transformations or randomization to the input of the model	Remove perturbations in adversarial examples	Simple; Do not modify the model	Attackers can bypass these defenses
	Feature squeezing [106]	Compare the predictions of the model on the original input and the squeezed input	Can detect adversarial examples	Simple, low computational overhead	Cannot defend against adaptive attackers
	MagNet [107]; Detector subnetwork [108]	Use detector network(s) to detect adversarial examples	Can detect (and reform [107]) adversarial examples	Generic to most NNs	Require training an additional detector network

在[94]中，Chen 等人提出了一种称为 SecDefender 的恶意软件检测方法，该方法利用分类器重新训练和安全正则化来增强分类器的鲁棒性。实验结果表明，即使攻击者知道目标分类器，SecDefender 也能有效防御规避攻击[94]。

Zhang 等人[95]提出了一种对抗特征选择方法来防御规避攻击。通过这种方法选择的特征可以提高模型的性能，也可以提高模型抵抗规避攻击的能力。对垃圾邮件和 PDF 恶意软件检测的评估结果表明，对抗性特征选择方法优于传统的特征选择方法[95]。在[96]中，Chen 等人选择攻击者难以操纵的特征来生成安全模型。此外，它们结合了多个模型的结果，以确保特征选择方法不会影响分类器的性能。类似地，Liang 等人[97]通过提取除非危及恶意软件的恶意功能否则不能被修改的特征来训练安全模型。Bhagoji 等人[98]使用降维来保护模型免受规避攻击。然而，这种减少特征尺寸的方法也可能影响模型的性能。

这些防御方法[4]、[5]、[43]、[89]-[98]简单而有效。然而，大多数防御方法集中于二进制分类问题，例如垃圾邮件检测，恶意软件检测和 IDS。这些方法可能不适用于其他机器学习任务。

2) 在深度学习算法的背景下对对抗性例子的防御

深度学习算法背景下对抗对抗性例子的防御技术可以进一步分为两种类型[7]，完全防御技术和仅检测技术。完整的防御技术旨在使模型能够识别敌对示例的正确标签。另一方面，仅检测技术只需要识别输入实例是否是潜在的敌对实例。

完全防御方法：一个普遍的结论是，对抗性训练可以作为对抗对抗性攻击的第一道防线[2]、[7]、[99]。Goodfellow 等人[2]提出了一种快速对抗样本生成方法，称为“快速梯度符号方法”(FGSM)，并建议使用这些生成的对抗样本进行对抗训练。研究表明，对抗性训练可以正则化模型，从而提高模型的鲁棒性[2]。Kurakin 等人[99]后来将上述对抗性训练方法扩展到大型模型和大型数据集。对抗训练简单，能有效提高目标模型的鲁棒性。然而，对抗性训练不可避免地增加了训练数据量和训练开销。此外，攻击者可以基于已经存在的模型再次生成对抗性的例子

进行了对抗性训练，这将演变成一场训练竞赛[7]。

Madry 等人[100]通过使用鞍点公式提出了一种针对对抗性攻击的鲁棒深度学习策略。他们还建议，模型容量对稳健性具有显著影响，因此需要更大的模型容量来抵抗对抗性攻击。顾和 Rigazio [101]提出了一个防御模型，命名为深度收缩网络(DCN)。DCN 在训练阶段使用类似于收缩自动编码器的平滑度惩罚。结果，它可以有效地使模型的输出对输入不太敏感[101]，从而可以增加模型对对立例子的鲁棒性。Papernot 等人[102]提出了一种防御技术，称为防御蒸馏，它使用两个网络，一个初始网络和一个蒸馏网络。概率向量中包含的知识通过蒸馏进行传递，因此可以提高 DNNs 的泛化能力，从而增强其对扰动的鲁棒性[102]。罗斯和多希-维勒兹[103]提出了一种通过调整 DNNs 输入梯度的防御技术。该方法将输出的变化程度对输入的变化进行惩罚。因此，小的对抗性扰动不能显著改变模型的预测。这些方法[100]-[103]是大多数神经网络的通用方法，但这些方法可能会增加模型的训练复杂性，或者可能会被更强的攻击方法击败[7]。例如，证明了防御蒸馏法[102]可以被卡里尼和瓦格纳攻击[49]击败。

Guo 等人[104]使用输入变换来对抗敌对图像。在将输入图像输入 ImageNet 之前，他们对其进行了转换，例如 JPEG 压缩、位深度缩减、图像拼接和总方差最小化。他们的实验结果表明，在对这些变换图像进行训练后，卷积网络分类器可以有效地防御敌对扰动。这表明一些图像变换操作可以消除扰动[104]。同样，谢等人[105]证明了在测试阶段的随机化可以减轻对立例子的影响。具体来说，在输入到模型中之前，对输入应用两种随机化操作，随机调整大小和随机填充。研究表明，这些随机化操作可以防御单步和迭代对抗性攻击[105]。这些方法[104]、[105]是简单的，并且不修改或重新训练目标模型。然而，攻击者也可以对生成的对抗性示例进行图像变换，使其具有鲁棒性，从而确保生成的对抗性示例能够绕过这些防御方法。

*仅检测方法:Xu 等人[106]使用特征压缩操作来检测对立示例。使用两种特征压缩操作，包括减少每个像素的颜色深度和执行空间平滑。之后，他们比较 DNN 模型对原始输入和压缩输入的预测[106]。如果预测之间存在较大差异，则输入为*被认为是对立的例子。该方法简单且计算开销低，但该方法无法防御完全了解防御技术的自适应攻击者[106]。孟和陈[107]亲提出一个防御框架，命名为磁铁。在 Mag- Net 中使用一个或多个检测器网络和一个重整器网络。被发现具有大扰动的实例被拒绝。另一方面，具有小扰动的实例被重整器网络[107]向正常实例重整，然后重整的实例被馈送到分类器中用于正确分类。Metzen 等人[108]通过增加一个检测器子网络来强化 DNN 模型，该子网络在识别输入中的敌对扰动的分类任务上受到训练。已经证明，将这样的检测器子网络添加到主分类网络可以有效地检测对立的例子[108]。虽然检测器只被训练来检测一种类型的攻击，但是它可以很好地推广到类似的或较弱的攻击。这些方法[107]、[108]使用检测器来确定目标模型的输入是否是对立示例，这些方法独立于目标模型，并且可以为不同的模型提供对立示例检测。然而，这些检测方法[107]、[108]需要训练额外的检测器来检测模型的输入，这通常会引入高计算开销。

D. 对模型窃取攻击的防御

在这一节中，我们将讨论如何防御模型窃取攻击。针对模型窃取攻击的防御措施总结如表所示 9。针对模型窃取的直观防御方法是，当输出标签时，不提供置信度信息。但是，这会影响服务质量。Lee 等人[109]通过在置信度信息中注入欺骗性扰动/噪声来误导对手，从而保护机器学习模型。结果，对手只能使用标签来窃取模型，并需要更多的查询来窃取模型[109]。这种方法在不影响模型精度的情况下保护了模型。但是，如果攻击者增加查询次数，攻击仍然可以成功。

Hanzlik 等人[110]提出了一个离线机器学习即服务框架，命名为 MLCapsule。在 MLCapsule 框架中，模型被允许在用户侧执行，使得用户可以保护他们的数据的隐私，同时服务器仍然可以控制具有知识产权保护的模型。然而，这种方法需要加密用户的数据，这将引入额外的计算开销。尤蒂等人[111]提出了一种检测 DNN 模型窃取攻击的技术，命名为 PRADA。由于恶意软件通过预测 API 窃取模型，PRADA 分析查询的分布并检测异常的连续查询。PRADA 是通用的，可以检测大多数模型的模型提取攻击。然而，这种方法对于用来隐藏攻击者的异常查询的伪查询[112]并不鲁棒。

表 9。模型窃取攻击防御综述。

Approaches	Working mechanism	Effect	Advantages	Disadvantages
Deceptive perturbations [109]	Injecting deceptive perturbations in the confidence information	Makes the confidence information useless for attackers	Protect the model without affecting the model accuracy	Attackers can still succeed with increased number of queries
MLCapsule [110]	Execute the model on the isolated execution environments	Provide a secure environment to run models	Protect both the model security and the users' data privacy	Increase computational and hardware overhead
PRADA [111]	Monitor the distribution of continuous queries	Detect anomalous queries	Generic	Not robust to dummy queries

E. 防止敏感训练数据恢复的隐私保护机器学习技术

机器学习模型对敏感训练数据恢复的防御可以大致分为三类:(1)基于密码原语的方法，例如差分隐私、同态加密；(2)分布式学习的安全聚集/集合，例如联合学习[113]，教师集合的私人聚集(PATE)[114]；(3)基于可信平台/处理器的方法。针对敏感训练数据恢复的隐私保护机器学习技术的总结在表中给出 10。

1) 基于密码原语的方法

Abadi 等人[115]开发了基于差分隐私的深度学习框架。此外，他们提出了提高基于差分隐私的训练的效率的技术，以便实现隐私、效率、软件复杂性和模型质量之间的平衡。基于不同隐私的方法[115]对大多数机器学习算法是通用的，但是该方法在训练模型的过程中向梯度添加了噪声会影响模型的准确性。Jayaraman 等人。[123]证明在当前的隐私保护机制中，模型的隐私和性能之间存在权衡。换句话说，当保护模型的隐私时，当前的隐私保护机制也将牺牲模型的性能[123]。Phong 等人[116]表明[74]中用于隐私保护的分布式学习框架仍然可能向服务器泄露秘密数据。因此，他们通过对 NN 使用异步随机梯度下降来改进[74]中的技术，并将同态加密引入框架[116]。基于同态加密的方法[116]使用密码原语来保证训练数据的安全性和私密性，而不影响模型的准确性，但这种方法不可避免地在模型的训练过程中带来较高的计算开销。Rouhani 等人[117]提出了一个可证明安全的学习框架，命名为 DeepSecure，其中使用了 Yao 的 gar- bled 电路协议来执行安全计算。他们还提出了优化实现和低开销技术来减少开销。虽然该方法比基于同态加密的方法更有效，但在实际应用中不容易部署。

表 10。防止敏感训练数据恢复的隐私保护机器学习技术综述。

Categories	Approaches	Working mechanism	Effect	Advantages	Disadvantages
Cryptographic primitive-based approaches	Deep learning with differential privacy [115]	Differential privacy-based	Prevent attackers from stealing the privacy of the training data	Generic	Affect the accuracy of the model
	Deep learning via additively homomorphic encryption [116]	Homomorphic encryption-based	Protect the gradients	Without affecting the accuracy of the model	High computational overhead
	DeepSecure [117]	Yao's garbled circuit and low-overhead techniques	Secure deep learning with low overhead	Efficient and scalable	Not easy to deploy in practice
Secure aggregation/ensembles of distributed learning	Distributed learning [74]; SecureML [118]	Jointly learn the model using asynchronously stochastic gradient descent [74]; Secure two-party computation [118]	Multiple users jointly learn a model without sharing their data	Protect the privacy of data for multiple users while achieve high model accuracy	High computational overhead
	Federated Learning [113]	Secure multi-party computation-based			
	PATE [114]; Scalable PATE [119]	Private aggregation of teacher ensembles	Prevent attackers from accessing data	Generic	Need to train an additional student model
Trusted platforms or processors-based approaches	Oblivious multi-party learning on trusted processors [120]; Deep learning on multi-source private data [121]	Run the model on trusted platforms	Prevent attackers from accessing the model	Low computational overhead	Require the support of hardware platforms
Specific application-oriented	Disguised-Nets [122]	Use image disguising techniques	Protect model and data privacy	Simple while maintaining good model performance	Only suitable for image data

2) 分布式学习的安全聚合/集成

分布式学习框架:Shokri 和 Shmatikov [74]提出了一种用于隐私保护的分布式学习框架，其中多个实体可以联合学习神经网络模型，而无需共享它们自己的数据，并且只共享学习参数的一个小子集。关键思想是基于随机梯度下降的深度学习算法可以并行执行[74]。Mohassel 和 Zhang [118]提出了保护隐私的机器学习模型的新协议，包括逻辑回归、线性回归和神经网络。该协议在两个服务器模型中工作，其中两个服务器通过使用安全的双方计算在分布式私有数据上训练它们自己的模型[118]。

联合学习的安全聚合:Bonawitz 等人[113]提出了一个基于安全多方计算(安全聚合)的联合学习框架。在分布式学习模型中，安全聚合可以保护每个用户模型的梯度信息。

教师群体的私人聚集:Papernot 等人[114]提出了一个名为 PATE 的隐私保护培训模型。他们使用不相交的敏感数据集训练了多个模型，称为教师模型。基于所有教师的嘈杂集合的输出来学习学生模型，因此不能访问单个教师模型的数据或参数[114]。Papernot 等人[119]将 PATE 扩展到大规模任务和带有错误的未切割数据集。具体来说，他们开发了新的噪声聚合方法，以较少的噪声集成教师模型。

分布式学习框架方法[74]、[118]和安全聚合方法[113]可以在分布式深度学习框架中保护多个用户的数据隐私，但在聚合多个用户的数据时使用的加密算法或安全协议会增加计算开销。PATE 方法[114]，[119]可以为大多数模型提供训练数据保护，但是 PATE 方法需要训练一个额外的学生模型，以便用户访问模型。

3) 基于可信平台/处理器的方法

另一个理想的隐私保护解决方案是在可信平台上运行机器学习模型。Ohrimenko 等人[120]在可信处理器英特尔 SGX 上开发机器学习模型。他们提出了用于 SVM、决策树、矩阵分解、k 均值聚类和神经网络的不经意机器学习算法。他们证明，这种基于可信处理器的机器学习实现比基于密码的隐私保护解决方案具有更高的性能[120]。Hynes 等人[121]提出了一个隐私保护的深度学习框架，命名为髓磷脂。髓磷脂结合了差分隐私和可信硬件飞地来训练机器学习模型。这些方法[120]、[121]使用可信平台防止攻击者访问模型，并为模型的训练数据提供安全的隐私保护机制。它们的计算开销低于基于加密的防御方法。然而，这些方法[120]、[121]依赖于可信平台，这需要硬件平台的支持。

4) 面向特定应用

Sharma 和 Chen [122]在基于图像的深度学习任务中使用图像伪装技术来保护隐私。他们证明了经过变换和分块排列的图像可以保证 DNN 模型的私密性和可用性。这种方法简单有效。但是，这种方法只适用于保护图像数据的隐私。

F. 关于防御的讨论

现有的保护方法可以总结如下。在训练阶段，针对中毒攻击或后门攻击的防御工作可以被称为数据净化[24]，其中异常中毒数据在进入训练阶段之前首先被过滤掉。异常检测器通常基于训练损失、最近邻居等[24]。

在测试阶段，针对不利例子的防御技术可以称为平滑模型输出[11]，即降低模型输出对输入变化的敏感性。针对敏感信息泄漏的防御技术包括三个主要类别:分布式学习框架、传统的基于密码原语的方法(例如，基于差分密码、同态加密)和基于可信平台的方法。

V. 安全评估

A. 安全设计

在典型的机器学习系统设计流程中，设计者关注的是模型选择和性能评估，而没有考虑安全问题。随着上述对机器学习系统的安全攻击的出现，有必要在设计阶段对机器学习系统进行安全评估，并使用最新的安全机器学习技术。这种范式可以称为安全设计，是对典型的性能设计范式的必要补充。例如，Biggio 等人[41]提出了分类器安全性评估的框架。它们通过增加对手的能力和知识来模拟不同层次的攻击。类似地，Biggio 等人[56]建议通过经验评估在一系列潜在攻击下的性能下降来评估分类器的安全性。特别是，它们生成训练集和测试集，并模拟攻击以进行安全评估。

B. 使用强攻击的评估

Carlini 和 Wagner [50]评估了十种最新的检测方法，并表明这些防御措施可以通过使用

具有新损失函数的强攻击。因此，建议使用强攻击对机器学习算法进行安全性评估，包括以下两个方面。首先，在白盒攻击下进行评估，例如，攻击者具有关于模型、数据和防御技术的完美知识，并且具有操纵数据或模型的强大能力。第二，在高置信度攻击/最大置信度攻击下进行评估，而不仅仅是最小扰动攻击[3]。Carlini 和 Wagner [49]，[50]表明，针对最小扰动攻击提出的防御技术可以通过使用高置信度攻击来绕过。关于对立例子的最初工作旨在分析深度学习算法对最小扰动的敏感性。然而，为了分析深度学习算法的安全性，使用最大置信度对抗性攻击更为合理，它可以反映算法在更强大攻击下的安全性[3]。

C. 评估指标

首先，建议使用更多的度量标准[50]，例如，不仅是准确性，而且还有混淆矩阵(真阳性、假阳性、真阴性、假阴性)、精确度、召回率、ROC(接收器操作特征)曲线和 AUC(ROC 曲线下的面积)，来报告学习算法的性能，从而可以反映完整的性能信息，并且易于与其他工作进行比较。第二，可以使用安全性评估曲线[3]。比格吉奥和花小蕾[3]提出使用安全评估曲线来评估学习系统的安全性。安全评估曲线表征了不同攻击强度和攻击者不同知识水平下的系统性能[3]，从而可以提供攻击下系统性能的综合评估，也便于比较不同的防御技术。

VI. 未来方向

机器学习安全是一个非常活跃的研究方向。这几年有很多针锋相对的攻击和防御的作品。我们提出了机器学习安全性的以下未来方向:

1) *真实物理条件下的攻击。针对机器学习模型的安全攻击有很多，其中大部分都在数字仿真实验中得到了验证。这些攻击在真实世界物理条件下的有效性，以及针对真实世界物理条件的工作，是活跃的研究主题。例如，物理对抗性例子可以欺骗路标识别系统，但是这些物理对抗性例子在视觉上是明显的和不自然的。最近，大量的工作旨在生成自然健壮的物理对抗样本。此外，基于 DNN 的智能监控系统已被广泛部署。对于人类来说，有可能在物体面前实现隐形吗通过对立例子的探测器？由于人类的大的类内差异，以及人类的动态运动和不同姿势，这是比数字对抗示例攻击和面向路标的对抗示例攻击更具挑战性的任务。*
2) *保护隐私的机器学习技术。近年来，机器学习的隐私问题受到越来越多的关注。深度学习的部署需要解决隐私保护的问题，包括从服务提供商的角度保护模型的参数，从用户的角度保护用户的隐私数据。迄今为止，基于密码原语的机器学习方法的效率需要提高，这通常会给模型的训练带来高开销，并且可能会降低模型的性能。分布式或基于集成的训练框架仍然面临效率和性能问题。有必要研究安全高效的机器学习算法、模型和框架。结合硬件平台、软件和算法的协同设计来保护 DNN 的隐私是一个有前途的方向。*
3) *DNN 的知识产权保护。深度学习模型的训练需要海量的训练数据，以及大量的硬件资源来支持。训练过程通常需要几周或几个月。从这个意义上说，机器学习模型是模型提供者有价值的商业知识产权，因此需要得到保护。目前，只有少数基于机器学习模型的水印知识产权保护作品[124]。对于 DNN 来说，更有效和更安全的知识产权保护方法仍然是公开的问题。*
4) *远程或轻量级机器学习安全技术。机器学习将广泛用于分布式、远程或物联网场景中的平台。在这些资源受限的情况下，许多现有的安全技术都不适用。如何提供可靠有效的远程或轻量级机器学习安全技术是一个很有前景的研究方向。*
5) *系统机器学习安全评估方法。迄今为止，在机器学习安全评估方面做的工作很少。具体来说，没有全面的方法来评估模型的安全性和稳健性以及模型的训练数据和参数的安全性和私密性。也没有统一的方法和全面的指标来评估当前攻击和防御的性能。需要研究和建立一种涉及机器学习系统的安全性、健壮性、隐私性的系统评估方法，以及相应的评估指标。*
6) *这些对机器学习的攻击和防御背后的深层原因是什么？文献中有一些讨论，但仍缺乏共识。*

这些攻击背后的原因仍然是个悬而未决的问题。此外，模型的不透明性使得它目前缺乏对模型输出的解释。然而，在一些关键应用中，如医疗保健和银行，应用模型的可解释性是必需的[11]。

VII. 结论

基于机器学习的应用无处不在，然而机器学习系统在其生命周期中仍然面临着各种各样的安全威胁。机器学习安全是一个活跃的研究课题，也是一个公开的问题。本文针对机器学习系统的五种主要攻击类型及其相应对策，对机器学习系统的安全进行了全面的综述，涵盖了机器学习系统的整个生命周期。总的结论是，威胁是真实的，新的安全威胁不断出现。例如，研究表明，对立示例具有可转移性，这意味着对立示例可以在不同的机器学习模型之间很好地推广。研究表明，中毒的例子也可以很好地概括不同的学习模型。这种可转移性可以有效地用于在黑盒场景中发起攻击。由于机器学习模型无法解释的性质，这些攻击的本质原因，即，对抗性示例是模型的缺陷还是固有属性，需要进一步研究。本文有望为设计安全、健壮和私有的机器学习系统提供全面的指导。

参考

[1] C.塞格迪、w .扎伦巴、I .苏茨基弗、j .布鲁纳、d .埃汉、I. J .古德费尔-洛和r .弗格斯，“神经网络的迷人特性”，正在进行中。第二国际。糖膏剂学习。代表。，2014年4月，第1-10页。
[2] I. J. Goodfellow、J. Shlens和C. Szegedy，“解释和利用对立的例子”，正在进行中。里面的糖膏剂学习。陈述，2015年3月，第1 - 11页。
[3] B.比格吉奥和花小蕾，“野生模式:十年后，对抗性机器学习的崛起，”模式识别。，第84卷，第317 - 331页，2018年12月。
[4] 名词（noun 的缩写）n .达尔维、P. M .多明戈斯、毛萨姆、S. K .桑海和d .维尔马，“对抗性分类”，正在进行中。第10届美国计算机学会国际会议。糖膏剂知道了。Discov。数据最小值。，2004年8月，第99 - 108页。
[5] B.Nelson、M. Barreno、F. J. Chi、A. D. Joseph、B. I. P. Rubinstein、U. Saini、C. A. Sutton、J. D. Tygar和K. Xia，“利用机器学习颠覆你的垃圾邮件过滤器”，正在进行中。USENIX作坊大规模开采。紧急情况。威胁。，2008年4月，第1 - 9页。
[6] 米（meter 的缩写））Barreno，B. Nelson，A. D. Joseph 和J. D. Tygar，《机器学习的安全性》，Mach。学习。，第81卷，第2期，第121-148页，2010年11月。
[7] 名词（noun 的缩写）Akhtar 和A. S. Mian，“对计算机视觉中深度学习的对抗性攻击的威胁:一项调查”，IEEE Access，第6卷，第14410-14430页，2018年7月。
[8] X.袁，何培平，朱清泉，李，“对抗的例子:深度学习的攻击与防御”，IEEE Trans。神经网络。学习。系统。，第30卷，第9期，第2805 - 2824页，2019年9月。
[9] 米（meter 的缩写））S. Riazi 和F. Koushanfar，“保护隐私的深度学习和推理”，正在进行中。里面的糖膏剂计算机。-辅助设计ICCAD，2018年11月，第1-4页。
[10] Q. Liu，P. Li，W. Zhao，W. Cai，S. Yu 和V. C. M. Leung，“关于机器学习的安全威胁和防御技术的调查:数据驱动的观点”，IEEE Access，第6卷，第12103-12117页，2018年。
[11] 名词（noun 的缩写）Papernot，P. D. McDaniel，A. Sinha 和M. P. Wellman，“SoK:机器学习中的安全和隐私”，正在进行中。IEEE Eur。症状。安全。隐私(EuroS&P)，2018年4月，第399 - 414页。
[12] T.顾，b .多兰-加维特和s .加格，“坏网:识别机器学习模型供应链中的漏洞”，2017年，arXiv:1708.06733。【在线】。可用:http://arxiv.org/abs/1708.06733
[13] 南约姆、I·贾科姆利、m·弗雷德里克松和s·杰哈，“机器学习中的隐私风险:分析过度拟合的联系”，正在进行中。IEEE第31届计算机大会。安全。找到了。症状。(CSF)，2018年7月，第268 - 282页。
[14] B.I. P. Rubinstein、B. Nelson、L. Huang、A. D. Joseph、S. Lau、S. Rao、N. Taft和J. D. Tygar，“解毒剂:了解和防范异常探测器中毒”，正在进行中。第九届美国计算机学会信号通信会议。互联网措施。糖膏剂(IMC)，2009年11月，第1-14页。
[15] 页（page 的缩写）李，刘，赵，王，王，“慢性中毒对基于机器学习的智能决策支持系统使用边缘模式检测，”在进步。IEEE Int。糖膏剂Commun。(国际商会)，2018年5月，第1 - 7页。
[16] B.比格吉奥、g .富梅拉、f .花小蕾和l .迪达奇，“中毒适应性生物计量系统”，正在进行中。里面的车间统计。技术。模式识别。结构。合成。模式识别。，2012年11月，第417 - 425页。
[17] B.比格吉奥、l .迪达奇、g .富梅拉和f .花小蕾，“对compromise人脸模板的中毒攻击”，正在进行中。里面的糖膏剂生物统计学(ICB)，2013年6月，第1 - 7页。
[18] B.Biggio，B. Nelson和P. Laskov，“针对支持向量机的中毒攻击”，在Proc。第29国际。糖膏剂马赫。学习。，2012年6月，第1467 - 1474页。
[19] B.比格乔，皮莱，S. R .布洛，d .艾利乌，m .佩利洛和f .花小蕾，“敌对环境下的数据聚类安全吗？’’进行中。ACM研讨会Artif。智能。安全。AISec，2013年11月，第87 - 98页。
[20] B.Biggio，K. Rieck，D. Ariu，C. Wressnegger，I. Corona，G. Giacinto和
F.花小蕾，“毒害行为恶意软件集群”，在Proc。Artif车间。智能。安全。AISec研讨会，2014年11月，第27 - 36页。
[21] H.肖，b .比吉奥，g .布朗，g .富梅拉，和f .花小蕾，“特征选择对训练数据中毒安全吗？”。第32国际。糖膏剂马赫。学习。，2015年7月，第1689 - 1698页。
[22] B.李，王，a .辛格，和y .沃罗比奇克，“基于因子分解的协同过滤的数据中毒攻击”，正在进行中。Annu Conf。神经感染。继续。系统。，2016年12月，第1885 - 1893页。
[23] 米（meter 的缩写））Jagielski，A. Oprea，B. Biggio，C. Liu，C. Nita-Rotaru和B. Li，“操纵机器学习:回归学习的中毒攻击和对策”，正在进行中。IEEE Symp。安全。隐私(SP)，2018年5月，第19 - 35页。
[24] 页（page 的缩写）W. Koh，J. Steinhardt，和P. Liang，“更强的数据中毒攻击突破数据净化防御”，2018年，arXiv:1811.00741。【在线】。
可用:http://arxiv.org/abs/1811.00741
[25] Y.王和K. Chaudhuri，“针对在线学习的数据中毒攻击”，2018年，arXiv:1808.08994。【在线】。可用:http://arxiv.org/abs/1808.08994
[26] X.张，x .朱，L. Lessard，“在线数据中毒攻击”，2019，*arXiv:1903.01666*。【在线】。可用:http://arxiv.org/abs/1903.01666
[27] C.杨，吴，李，陈，“神经网络的生成性中毒攻击方法”，2017，arXiv:1703.01340 .【在线】。可用:http://arxiv.org/abs/1703.01340
[28] 长度穆尼奥斯·冈萨雷斯、比吉奥、德蒙蒂斯、波戴斯、
动词（verb 的缩写）旺格拉萨米，E. C .卢普和f .花小蕾，“反向梯度优化对深度学习算法的毒害”，正在进行中。第十届美国计算机学会研讨会。智能。安全。AISec，2017年，第27 - 38页。
[29] 米（meter 的缩写））莫扎法里-克马尼、s .苏尔科雷、a .拉古纳坦和N. K .贾，“医疗保健中机器学习的系统中毒攻击和防御”，IEEE J. Biomed。健康信息。，第19卷，第6期，第1893 - 1905页，2015年11月。
[30] 米（meter 的缩写））方，杨国荣，龚正志，刘军，“基于图的推荐系统的中毒攻击”，在会议录。第34届Annu。计算机。安全。应用Conf。，2018年12月，第381 - 392页。
[31] C.苗，李，肖，江，怀，苏，“群体感知系统中的数据中毒攻击”，正在进行中。第18届美国计算机学会国际。症状。移动自组织网络。计算机。- Mobihoc，2018年6月，第111 - 120页。
[32] Y.纪，张，王，“学习系统的后门攻击”，正在进行中。IEEE会议。Commun。Netw。安全。(CNS)，2017年10月，第1 - 9页。
[33] X.Chen，C. Liu，B. Li，K. Lu和D. Song，“利用数据中毒对深度学习系统进行有针对性的后门攻击”，2017年，arXiv:1712.05526。【在线】。可用:http://arxiv.org/

abs/1712.05526
[34] C.廖，钟，斯奎查里尼，朱，米勒，“基于隐形扰动的卷积神经网络模型的后门嵌入”，2018，arXiv:1808.10307。【在线】。可用:http://arxiv.org/abs/1808.10307

[35] 米（meter 的缩写））Barni，K. Kallas 和 B. Tondi，“通过培训设置腐败而不进行标签中毒的 CNN 新后门攻击”，2019 年，arXiv:1902.11237。【在线】。可用:http://arxiv.org/abs/1902.11237

[36] E.Bagdasaryan，A. Veit，Y. Hua，D. Estrin，V. Shmatikov，“如何借壳联邦学习”，2018，arXiv:1807.00459。【在线】。可用:http://arxiv.org/abs/1807.00459

[37] 米（meter 的缩写））邹，史耀辉，王春红，李芳辉，宋文辉，王耀辉，“深度学习模型中的强大神经级木马设计”，2018，arXiv:1802.03043。【在线】。可用:http://arxiv.org/abs/1802.03043

[38] Y.刘，马，杨，李，翟，王，张，“神经网络中的木马攻击”，正在进行中。Netw。Distrib。系统。安全。症状。，2018 年 2 月，第 1 - 15 页。

[39] Y.刘，谢，斯里瓦斯塔瓦，“神经木马”，在 Proc。IEEE Int。糖膏剂计算机。设计(ICCD)，2017 年 11 月，第 45 - 48 页。

[40] D.Lowd 和 C. Meek，“对抗性学习”，在进行中。第 11 届美国计算机学会国际会议。糖膏剂知道了。Discov。数据挖掘，2005 年 8 月，第 641 - 647 页。

[41] B.放大图片作者:Julian j .
G.贾辛托和 f .花小蕾，“在测试时规避对机器学习的攻击”，正在进行中。联合欧元。糖膏剂马赫。学习。知道了。发现数据库，2013 年 9 月，第 387 - 402 页。

[42] 名词（noun 的缩写）Rndic 和 P. Laskov，“基于学习的分类器的实际规避:案例研究”，正在进行中。IEEE Symp。安全。隐私，2014 年 5 月，第 197 - 211 页。

[43] A.德蒙蒂斯，m .梅利斯，b .比吉奥，d .马约尔卡，D. Arp，k .里克，
I. 科罗纳、g·贾辛托和 f·花小蕾，“是的，机器学习可以更安全！Android 恶意软件检测案例研究。可靠的安全计算。，第 16 卷，第 4 期，第 711 - 724 页，2019 年 7 月。

[44] W.徐，Y. Qi 和 D. Evans，“自动回避分类器”，正在进行中。Netw。Distrib。系统。症状。，2016 年 2 月，第 1 - 15 页。

[45] H.党，y .黄，和 E. C .张，〈通过在黑暗中变形来逃避分类器〉。ACM SIGSAC 会议。计算机。Commun。安全。(CCS)，2017 年，第 119 - 133 页。

[46] A.Kurakin，I. J. Goodfellow 和 S. Bengio，“物理世界中的对立例子，”在 Proc。里面的糖膏剂学习。代表。(ICLR)研讨会，2017 年 2 月，第 1 - 14 页。

[47] 名词（noun 的缩写）Papernot，P. D. McDaniel，S. Jha，m .弗雷德里克松，Z. B .切利克和 A. Swami，“深度学习在对抗性环境中的局限性”，在 Proc。IEEE Eur。症状。安全。隐私(EuroS&P)，2016 年 3 月，第 372 - 387 页。

[48] J.苏，D. V. Vargas，K. Sakurai，“愚弄深度神经网络的一个像素攻击”，IEEE Trans。伊沃。计算机。，第 23 卷，第 5 期，第 828 - 841 页，2019 年 10 月。

[49] 名词（noun 的缩写）卡里尼和 D. A .瓦格纳，“评估神经网络的鲁棒性”，在 Proc。IEEE Symp。安全。隐私(SP)，2017 年 5 月，第 39 - 57 页。

[50] 名词（noun 的缩写）Carlini 和 D. A. Wagner，“对立的例子不容易被发现:绕过十种检测方法”。第十届 ACM 研讨会。智能。安全。(AISec)，2017 年，第 3 - 14 页。

[51] 南 Moosavi-Dezfooli，A. Fawzi，O. Fawzi 和 P. Frossard，“普遍对抗性干扰”，正在进行中。IEEE 会议。计算机。Vis。模式识别。(CVPR)，2017 年 7 月，第 86 - 94 页。

[52] K.Grosse，N. Papernot，P. Manoharan，M. Backes 和 P. D. McDaniel，“恶意软件检测的对抗实例”，正在进行中。第 22 欧元。症状。关于计算机。安全。，2017 年 9 月，第 62 - 79 页。

[53] 米（meter 的缩写））Sharif，S. Bhagavatula，L. Bauer 和 M. K. Reiter，“犯罪的附属品:对最先进的人脸识别技术的真实和秘密的攻击”，正在进行中。ACM SIGSAC 会议。计算机。Commun。安全。CCS，2016 年，第 1528 - 1540 页。

[54] K.埃霍尔特、埃夫季莫夫、费尔南德斯、李、、肖、
A.Prakash，T. Kohno 和 D. Song，“对深度学习视觉分类的强大物理世界攻击”，在 Proc。IEEE/CVF 会议。计算机。Vis。模式识别。，2018 年 6 月，第 1625 - 1634 页。

[55] A.阿萨里，l .英斯特朗姆，a .易勒雅斯和 k .郭，“综合强大的对抗性的例子，”在 Proc。第 35 国际。糖膏剂马赫。学习。，2018 年 7 月，第 284 - 293 页。

[56] B.比吉奥、g .富梅拉和 f .花小蕾，“模式分类器在攻击下的安全性评估”，IEEE Trans。知道了。数据工程。，第 26 卷，第 4 期，第 984 - 996 页，2014 年 4 月。

[57] 南 Moosavi-Dezfooli，A. Fawzi 和 P. Frossard，“DeepFool:欺骗深度神经网络的简单而精确的方法”，在 Proc。IEEE 会议。计算机。Vis。模式识别。(CVPR)，2016 年 6 月，第 2574 - 2582 页。

[58] 名词（noun 的缩写）Papernot，P. D. McDaniel，I. J. Goodfellow，S. Jha，Z. B .切利克和
A.Swami，“针对机器学习的实用黑盒攻击”，正在进行中。亚洲会议。计算机。Commun。安全。(亚洲 CCS)，2017 年，第 506 - 519 页。

[59] J.陆，H. Sibai，E. Fabry，D. A. Forsyth，“无需担心自动驾驶汽车中物体检测的对抗性示例”，2017 年，arXiv:1707.03501。【在线】。可用:http://arxiv.org/abs/1707.03501

[60] 名词（noun 的缩写）Papernot 等人，《关于 CleverHans v2.1.0 对抗范例库的技术报告》，2018，arXiv:1610.00768。【在线】。可用:http://arxiv.org/abs/1610.00768

[61] D.Goodman，X. Hao，Y. Wang，Y. Wu，J. Xiong，和 H. Zhang，“Advbox:生成欺骗神经网络的对抗示例的工具箱”，2020 arXiv:2001.05574。【在线】。可用:http://arxiv.org/abs/2001.05574

[62] A.P. Norton 和 Y. Qi，“对抗性游戏场:展示对抗性例子如何愚弄深度学习的可视化套件”，正在进行中。IEEE Symp。可视化。网络安全。(VizSec)，2017 年 10 月，第 1 - 4 页。

[63] F.Tramèr，F. Zhang，A. Juels，M. K. Reiter 和 T. Ristenpart，“通过预测 API 窃取机器学习模型”，在 Proc .第 25 届 USENIX 安全。症状。，2016 年 8 月，第 601 - 618 页。

[64] Y.Shi，Y. Sagduyu 和 A. Grushin，“如何利用深度学习窃取机器学习分类器”，在 Proc。IEEE Int。症状。技术。本土安全。(HST)，2017 年 4 月，第 1 - 5 页。

[65] 动词（verb 的缩写）Chandrasekaran，K. Chaudhuri，I .，S. Jha 和 S. Yan，“探索主动学习和模型提取之间的联系”，2018 年，arXiv:1811.02054。【在线】。可用:http://arxiv.org/abs/1811.02054

[66] B.王和龚正志，“机器学习中的超参数窃取”，正在进行中。IEEE Symp。安全。隐私(SP)，2018 年 5 月，第 36 - 52 页。

[67] 南 Milli，L. Schmidt，A. D. Dragan 和 M. Hardt，“从模型解释重建模型”，正在进行中。糖膏剂公平、问责、透明 FAT，2019 年 1 月，第 1 - 9 页。

[68] 米（meter 的缩写））弗雷德里克松、e .兰茨、S. Jha、S. Lin、D. Page 和 T. Ristenpart，“药物遗传学中的隐私:个体化华法林给药的端到端案例研究”，正在进行中。第 23 届 USENIX 安全。症状。，2014 年 8 月，第 17 - 32 页。

[69] 米（meter 的缩写））弗雷德里克松、S. Jha 和 T. Ristenpart，“利用置信信息和基本对策的模型反演攻击”，在 Proc。第 22 届美国计算机学会会议。计算机。Commun。安全。CCS，2015 年，第 1322-1333 页。

[70] R.Shokri，M. Stronati，C. Song 和 V. Shmatikov，“对机器学习模型的成员推理攻击”，正在进行中。IEEE Symp。安全。隐私(SP)，2017 年 5 月，第 3 - 18 页。

[71] B.Hitaj，G. Ateniese 和 F. Perez-Cruz，“GAN 下的深度模型:合作深度学习的信息泄漏”，正在进行中。ACM SIGSAC 会议。计算机。Commun。安全。CCS，2017 年，第 603 - 618 页。

[72] 南 Hidano，T. Murakami，S. Katsumata，S. Kiyomoto 和 G. Hanaoka，“预测系统的模型反演攻击:没有非敏感属性的知识”，正在进行中。15 年。糖膏剂隐私安全。信托(PST)，2017 年 8 月，第 115 - 126 页。

[73] B.刘，吴，h .李，y .陈，q .吴，M. Barnell 和 q .邱，“克隆你的思维:认知系统设计中的安全挑战及其解决方案”，正在进行中。第 52 届 Annu。设计自动化。糖膏剂- DAC，2015 年 6 月，第 1-5 页。

[74] R.Shokri 和 V. Shmatikov，“保护隐私的深度学习”，正在进行中。第 53 届 Annu。阿勒顿会议。Commun。，控制，计算。艾伦顿出版社，2015 年 9 月，第 1310-1321 页。

[75] 南 Chen，C. Cornelius，J. Martin 和 D. H. Chau，“变形人:对更快的 R-CNN 物体探测器的强大物理对抗攻击”，正在进行中。联合欧元。糖膏剂马赫。学习。知道了。探索数据库，2018 年 9 月，第 52 - 68 页。

[76] 南 Thys，W. V. Ranst 和 T. Goedeme，“愚弄自动监视摄像机:攻击人员检测的对抗性补丁”，正在进行中。IEEE/CVF 会议。计算机。Vis。模式识别。研讨会(CVPRW)，2019 年 6 月，第 1 - 7 页。

[77] B.Biggio，I. Corona，G. Fumera，G. Giacinto 和 f .花小蕾，“对抗分类任务中对抗中毒攻击的 Bagging 分类器”，第 10 届国际会议录。糖膏剂 Mult。Classif。系统。，2011 年 6 月，第 350 - 359 页。

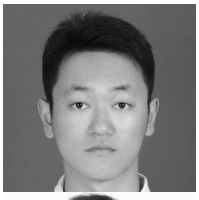
[78] 南陈，m .薛，l .范，s .郝，l .徐，h .朱，和 b .李，“自动投毒攻击和防御的恶意软件检测系统:一个对抗性的机器学习方法，”计算机。安全。，第 73 卷，第 326 - 344 页，2018 年 3 月。

[79] R.张和朱青，“分布式支持向量机中数据中毒攻击的博弈论防御”，在 Proc。IEEE 第 56 届年会。糖膏剂十分钟。控制(疾病预防控制中心)，2017 年 12 月，第 4582 - 4587 页。

[80] C.刘，b .李，Y. Vorobeychik 和 A. Oprea，“针对训练数据中毒的稳健线性回归”，正在进行中。第十届美国计算机学会研讨会。智能。安全。AISec，2017 年 11 月，第 91 – 102 页。
[81] J.Steinhardt，P. W. Koh 和 P. S. Liang，“数据中毒攻击的认证防御”，正在进行中。安奴。糖膏剂神经感染。过程。系统。，2017 年 12 月，第 3517 – 3529 页。
[82] 名词（noun 的缩写）Baracaldo，B. Chen，H. Ludwig 和 J. A. Safavi，“减轻对机器学习模型的 poi- soning 攻击:一种基于数据来源的方法”，正在进行中。第十届美国计算机学会研讨会。智能。安全。AISec，2017 年，第 103 – 110 页。
[83] 南沈，S. Tople 和 P. Saxena，“AUROR:在协作深度学习系统中防御中毒攻击”，正在进行中。第 32 届 Annu。糖膏剂计算机。安全。应用程序(ACSAC)，2016 年，第 508 – 519 页。
[84] B.Chen，W. Carvalho，N. Baracaldo，H. Ludwig，B. Edwards，T. Lee，I. Molloy 和 B. Srivastava，“通过激活聚类检测对深度神经网络的后门攻击”，在《美国科学学报》上发表。AAAI 工作室 Artif。智能。Saf。，2019 年 1 月，第 66 – 73 页。
[85] K.刘，b .多兰-加维特和 s .加格，“精细修剪:防御对深层神经网络的后门攻击”，正在进行中。第 21 国际。症状。攻击入侵防御。，2018 年 9 月，第 273 – 294 页。
[86] B.王，姚，山，李，李，郑，赵，“神经净化:识别和减轻神经网络中后门攻击”，正在进行中。IEEE Symp。安全。隐私(SP)，2019 年 5 月，第 707 – 723 页。
[87] Y.高，徐，王，陈，拉那辛和尼帕尔，“STRIP:深层神经网络对特洛伊木马攻击的防御”，正在进行中。第 35 届 Annu。计算机。安全。应用 Conf。ACSAC，2019 年，第 113 – 125 页。
[88] B.Gia Doan，E. Abbasnejad，D. C. Ranasinghe，“二月:深度神经网络系统上针对木马攻击的输入净化防御”，2019 年，arXiv:1908.03369。【在线】。可用:http://arxiv.org/abs/1908.03369
[89] A.kocz 和 C. H. Teo，“提高分类器鲁棒性的特征加权”，在 Proc .第六次会议。电子邮件反垃圾邮件，2009 年 7 月，第 1 – 8 页。
[90] A.Globerson 和 S. T. Roweis，“测试时的噩梦:通过特征删除的稳健学习”，在 Proc。第 23rd Int。糖膏剂马赫。学习。(ICML)，2006 年 6 月，第 353-360 页。
[91] B.比格吉奥、g .富梅拉和 f .花小蕾，“敌对环境中稳健分类器设计的多分类器系统”。j .马赫。学习。赛博恩。，第 1 卷，第 1-4 期，第 27-41 页，2010 年 12 月。
[92] 名词（noun 的缩写）rndi 和 P. Laskov，“基于分层文档结构的恶意 PDF 文件检测”，正在进行中。20 年。Netw。Distrib。系统。安全。症状。，2013 年 2 月，第 1 – 16 页。
[93] B.李和 Y. Vorobeychik，“二元域上的规避-稳健分类”，美国计算机学会会刊。知道了。《从数据中发现》，第 12 卷第 4 期，第 1 – 32 页，2018 年 7 月。
[94] 长度 Chen，Y. Ye 和 T. Bourlai，“恶意软件检测中的对抗性机器学习:规避攻击和防御之间的军备竞赛”，正在进行中。欧元。智能。安全。信息技术。糖膏剂(EISIC)，2017 年 9 月，第 99 – 106 页。
[95] F.张，陈佩刚，比吉奥，杨德生，，“对抗规避攻击的对抗特征选择”，《电气工程学报》。赛博恩。，第 46 卷，第 3 期，第 766 – 777 页，2016 年 3 月。
[96] 长度 Chen，S. Hou，和 Y. Ye，“SecureDroid:增强基于机器学习的检测的安全性，以对抗对抗性 Android 恶意软件攻击”，正在进行中。第 33 届 Annu。计算机。安全。应用 Conf。ACSAC，2017 年，第 362 – 372 页。
[97] 长度 Tong，B. Li，C. Hajaj，C. Xiao，N. Zhang 和 Y. Vorobeychik，“使用保守特征提高最大似然分类器抵抗可实现规避攻击的鲁棒性”，在《中国科学报》上发表。第 28 届 USENIX 安全。症状。，2019 年，第 285 – 302 页。
[98] A.N. Bhagoji，D. Cullina，C. Sitawarin 和 P. Mittal，“通过数据转换增强机器学习系统的健壮性”，正在进行中。第 52 届 Annu。糖膏剂 Inf。Sci。系统。(CISS)，2018 年 3 月，第 1 – 5 页。
[99] A.Kurakin，I. J. Goodfellow 和 S. Bengio，“大规模对抗性机器学习”，正在进行中。第五国际。糖膏剂学习。代表。，2017 年 4 月，第 1 – 17 页。
[100] A.Madry，A. Makelov，L. Schmidt，D. Tsipras 和 A. Vladu，“走向抵抗对抗性攻击的深度学习模型”，正在进行中。第六国际。糖膏剂学习。代表。，2018 年 5 月，第 1 – 10 页。
[101] 南 Gu 和 L. Rigazio，“走向对对立实例稳健的深度神经网络结构”，在《进展。第三国际。糖膏剂学习。代表。，2015 年 5 月，第 1 – 9 页。
[102] 名词（noun 的缩写）Papernot，P. McDaniel，X. Wu，S. Jha 和 A. Swami，“蒸馏作为对抗深度神经网络的对抗性扰动的防御”，正在进行中。IEEE Symp。安全。隐私(SP)，2016 年 5 月，第 582 – 597 页。

[103] A.S. Ross 和 f . Doshi-维勒兹，“通过调整输入梯度来提高深度神经网络的对抗性鲁棒性和可解释性”，在 Proc。第三十二届 AAAI 会议。阿提夫。英特尔。，2018 年 2 月，第 1660 - 1669 页。
[104] C.郭、m .拉纳、m .西塞和 l .范德马腾，“使用输入转换对抗敌对图像”，正在进行中。第六国际。糖膏剂学习。代表。，2018 年 5 月，第 1 - 12 页。
[105] C.谢，j .王，z .张，z .任和 A. L .尤耶，“通过随机化减轻对抗效应”，正在进行中。第六国际。糖膏剂学习。代表。，2018 年 5 月，第 17 - 32 页。
[106] W.徐，d .埃文斯，和 y .齐，“特征压缩:检测深层神经网络中的对立例子”，在会议录。Netw。Distrib。系统。安全。症状。，2018 年 2 月，第 1 - 15 页。
[107] D.孟和陈，“磁铁:对抗实例的双向防御”，正在进行中。ACM SIGSAC 会议。计算机。Commun。安全。CCS，2017 年，第 135-147 页。
[108] J.H. Metzen，T. Genewein，V. Fischer 和 B. Bischoff，“关于检测对抗性干扰”，正在进行中。里面的糖膏剂学习。代表。(ICLR)，2017 年 2 月，第 1-12 页。
[109] T.Lee，B. Edwards，I. Molloy 和 D. Su，“使用欺骗性扰动防御神经网络模型窃取攻击”，在《过程。IEEE 安全。隐私研讨会(SPW)，2019 年 5 月，第 43 - 49 页。
[110] 长度汉兹利克、张、格罗斯、萨利姆、奥古斯丁、巴克斯和米（meter 的缩写））Fritz，“MLCapsule:机器学习即服务的守护式离线部署”，2018，arXiv:1808.00590。【在线】。可用:http://arxiv。org/abs/1808.00590
[111] 米（meter 的缩写））尤蒂、s·西利勒、s·马沙尔和 n·阿索坎，“普拉达:防范 DNN 模特盗用攻击”，正在进行中。IEEE Eur。症状。安全。隐私(EuroS&P)，2019 年 6 月，第 512 - 527 页。
[112] 南陈，n .卡里尼，d .瓦格纳，“黑盒对抗性攻击的状态检测”，2019 年，arXiv:1907.05587。【在线】。可用:http://arxiv.org/abs/1907.05587
[113] K.博纳威兹、伊万诺夫、克罗伊特、马塞多内、麦克马汉、南 Patel，D. Ramage，A. Segal 和 K. Seth，“用户持有数据联邦学习的实用安全聚合”，2016 年，arXiv:1611.04482。【在线】。可用:http://arxiv.org/abs/1611.04482
[114] 名词（noun 的缩写）Papernot，M. Abadi，U. Erlingsson，I. J. Goodfellow 和 K. Talwar，“从私人训练数据进行深度学习的半监督知识转移”，正在进行中。第五国际。糖膏剂学习。代表。，2017 年 4 月，第 1 - 16 页。
[115] 米（meter 的缩写））阿巴迪，朱，古德菲勒，麦克马汉，米罗诺夫，
K.Talwar 和 L. Zhang，“具有不同隐私的深度学习”，正在进行中。ACM SIGSAC 会议。计算机。Commun。安全。CCS，2016 年 10 月，第 308 - 318 页。
[116] 长度冯国良，杨雅诺，T. Hayashi，L. Wang 和 S. Moriai，“通过附加同态加密保护隐私的深度学习”，IEEE Trans .Inf。《取证安全》，第 13 卷第 5 期，第 1333 - 1345 页，2018 年 5 月。
[117] B.d .鲁哈尼、M. S. Riazi 和 F. Koushanfar，“DeepSecure:可扩展的可证明安全的深度学习”，正在进行中。第 55 届美国计算机学会/ESDA/电气和电子工程师协会设计自动化。糖膏剂(DAC)，2018 年 6 月，第 1 - 6 页。
[118] 页（page 的缩写）Mohassel 和 Y. Zhang，“SecureML:一个可扩展的保护隐私的机器学习系统”，正在进行中。IEEE Symp。安全。隐私(SP)，2017 年 5 月，第 19 - 38 页。
[119] 名词（noun 的缩写）Papernot，S. Song，I. Mironov，A. Raghunathan，K. Talwar 和
单位 Erlingsson，“PATE 的可扩展私人学习”，正在进行中。第六国际。糖膏剂学习。代表。，2018 年 5 月，第 1 - 34 页。
[120] O.Ohrimenko、F. Schuster、C. Fournet、A. Mehta、S. Nowozin、K. Vaswani 和 M. Costa，“可信处理器上的不经意多方机器学习”，正在进行中。第 25 届 USENIX 安全。症状。，2016 年 8 月，第 619 - 636 页。
[121] 名词（noun 的缩写）Hynes，R. Cheng 和 D. Song，“多源私人数据上的高效深度学习”，2018 年，arXiv:1807.06689。【在线】。可用:http://arxiv.org/abs/1807.06689
[122] 南 Sharma 和 K. Chen，“伪装网络:用于保护隐私的外包深度学习的图像伪装”，2019 年，arXiv:1902.01878。【在线】。可用:http://arxiv.org/abs/1902.01878
[123] B.Jayaraman 和 D. Evans，“在实践中评估不同的私人机器学习”，在 Proc。第 28 届 USENIX 安全。症状。，2019 年，第 1895 - 1912 页。
[124] Y.Adi、C. Baum、M. Cisse、B. Pinkas 和 J. Keshet，“将你的弱点转化为优势:通过后门为深度神经网络添加水印”，正在进行中。第 27 届 USENIX 安全会议。症状。，2018 年 8 月，第 1615 - 1631 页。

薛明福(IEEE 成员)于 2014 年获得中国南京东南大学信息与通信工程博士学位。2011 年 7 月至 2012 年 7 月，他是新加坡南洋理工大学的研究实习生。他现在是南京航空航天大学计算机科学与技术学院的助理教授。从 2014 年到 2019 年，他主持了十个研究项目
相关期刊和国际会议上发表了大约 30 篇文章。件安全性、硬件特洛伊检测、人工智能安全性以习系统。他是 ACM、IEICE、CCF、CAAI 会员，中业委员会委员，ACM 南京分会执行委员，江苏省计算机网络与分布式计算专业委员会委员。他还是十多次国际会议的技术项目委员会成员。他获得了 ICCCS2015 最佳论文奖。他是第三届中国硬件安全研讨会的项目主席。



袁于 2017 年获得盐城师范学院数字媒体技术学士学位。他目前在中国南京的南京航空航天大学计算机科学与技术学院攻读硕士学位。他的研究兴趣包括人工智能安全、安全和私有机器学习系统。

吴合义于 2013 年获得东南大学信息工程硕士学位。他于 2013 年加入苏州国家税务局，担任网络安全工程师。他现任南京优普赛网络安全技术研究院有限公司的首席技术官。他还是两个安全组织(FreeBuf & Pediy)的讲师。他的研究兴趣包括网络安全、人工智能安全，
他是中国人工智能协会(CAAI)人工智能与安全专

张玉书(IEEE 成员)于 2014 年 12 月从中国重庆市重庆大学计算机科学学院获得博士学位。他曾在香港城市大学、西南大学、澳门大学和迪肯大学担任不同的研究职位。他现在是中国南京航空航天大学计算机科学与技术学院的教授。他发表了 100 多篇评论这些领域的期刊文章和会议论文。他的研究兴趣包括多媒体安全、人工智能、云计算安全、大数据安全、物联网安全和区块链。他是《信号处理》的编辑。

刘伟强(IEEE 资深会员)于 2006 年获得中国南京南京航空航天大学(NUAA)信息工程学士学位，并于 2012 年获得英国贝尔法斯特女王大学(QUB)电子工程博士学位。2012 年 8 月至 2013 年 11 月，他是 QUB 电子、通信和信息技术研究所(ECIT)的研究员。2013 年 12 月，他加入
学院，现任副教授。他出版了一本 Artech House
篇领先的期刊文章和会议论文。他的一篇文章是
RANSACTIONS ON COMPUTERS (TC)期刊的专题论括近似计算、计算机算法、数字信号处理的硬件安全和 VLSI 设计以及密码学。他是多个国际会议的技术项目委员会成员，包括 ARITH、DATE、ASAP、ISCAS、ASP-DAC、ICONI、SiPS、ISVLSI 和 NANOARCH，以及 IEEE 多规模计算系统汇刊(TMSCS)的指导委员会成员。他的文章在 IEEE 国际电路与系统研讨会(ISCAS 2011)和 ACM GLSVLSI 2015 上获得最佳论文奖。他是《IEEE 计算机汇刊》、《IEEE 电路和系统汇刊——I:常规论文》和《IEEE 计算新兴主题汇刊》的副主编，也是《IEEE 和 IEEE TETC 学报》(两个特刊)的客座编辑。他还是 IEEE ARITH 2020 的项目联合主席。