软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006131]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

机器学习模型安全与隐私研究综述*

纪守领[1], 杜天宇[1], 李进锋[1], 沈 超[2], 李 博[3]

[1](浙江大学网络空间安全研究中心 浙江大学计算机科学与技术学院,浙江 杭州 310027)

[2](西安交通大学智能网络与网络安全教育部重点实验室 西安交通大学电子与信息学部,陕西 西安 710049)

[3](伊利诺伊大学香槟分校计算机科学学院,伊利诺伊 厄巴纳香槟 61822, 美国)

通讯作者: 纪守领, E-mail: sji@zju.edu.cn

摘 要: 在大数据时代下,深度学习、强化学习以及分布式学习等理论和技术取得的突破性进展,为机器学习提供了数据和算法层面的强有力支撑,同时促进了机器学习的规模化和产业化发展.然而,尽管机器学习模型在现实应用中有着出色的表现,但其本身仍然面临着诸多的安全威胁.机器学习在数据层、模型层以及应用层面临的安全和隐私威胁呈现出多样性、隐蔽性和动态演化的特点.机器学习的安全和隐私问题吸引了学术界和工业界的广泛关注,一大批学者分别从攻击和防御的角度对模型的安全和隐私问题进行了深入的研究,并且提出了一系列的攻防方法.在本综述中,我们回顾了机器学习的安全和隐私问题,并对现有的研究工作进行了系统的总结和科学的归纳,同时明确了当前研究的优势和不足.最后,我们探讨了机器学习模型安全与隐私保护研究当前所面临的挑战以及未来潜在的研究方向,旨在为后续学者进一步推动机器学习模型安全与隐私保护研究的发展和应用提供指导.

关键词: 机器学习;投毒攻击;对抗样例;模型隐私;人工智能安全

中图法分类号: TP311

中文引用格式: 纪守领,杜天宇,李进锋,沈超,李博.机器学习模型安全与隐私研究综述.软件学报,2021. http://www.jos.org.cn/1000-9825/6131.htm

英文引用格式: Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: a survey. Ruan Jian Xue Bao/Journal of Software, 2021 (in Chinese). http://www.jos.org.cn/1000-9825/6131.htm

Security and Privacy of Machine Learning Models: A Survey

JI Shou-Ling[1], DU Tian-Yu[1], LI Jin-Feng[1], SHEN Chao[2], LI Bo[3]

[1](Institute of Cyberspace Research and College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

[2](Ministry of Education Key Laboratory for Intelligent Networks and Network Security and Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

[3](Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana-Champaign 61822, United States)

Abstract: In the era of big data, breakthroughs in theories and technologies of deep learning, reinforcement learning, and distributed learning have provided strong support for machine learning at the data and the algorithm level, as well as have promoted the development of scale and industrialization of machine learning. However, though machine learning models have excellent performance in many real-world applications, they still suffer many security and privacy threats at the data, model, and application levels, which could be characterized by diversity, concealment, and dynamic evolution. The security and privacy issues of machine learning have attracted extensive attention from academia and industry. A large number of researchers have conducted in-depth research on the security and privacy issues of models from the perspective of attack and defense, and proposed a series of attack and defense methods. In this survey,

* 基金项目: 国家重点研发计划项目(2018YFB0804102);浙江省自然科学基金杰出青年项目(LR19F020003);浙江省科技计划项目(2019C01055);国家自然科学基金项目(61772466,U1936215,U1836202,61822309,61773310,U1736205)

Foundation item: National Key Research and Development Program of China (2018YFB0804102); Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars (LR19F020003); Provincial Key Research and Development Program of Zhejiang, China (2019C01055); National Natural Science Foundation of China (61772466,U1936215,U1836202,61822309,61773310,U1736205)

收稿时间: 2019-06-10; 修改时间: 2019-10-15; 采用时间: 2020-08-17; jos 在线出版时间: 2020-09-10

we review the security and privacy issues of machine learning, systematically and scientifically summarize existing research work, and clarify the advantages and disadvantages of current research. Finally, we explore the current challenges and future research directions of machine learning model security and privacy research, aiming to provide guidance for follow-up researchers to further promote the development and application of machine learning model security and privacy research.

Key words: machine learning; poisoning attack; adversarial example; model privacy; artificial intelligence security

在大数据时代下,深度学习、强化学习以及分布式学习等理论和技术取得的突破性进展,为机器学习在计算机视觉、自然语言处理以及语音识别等多个领域的蓬勃发展提供了数据和算法层面的强有力支撑,同时也促进了机器学习技术在诸如自动驾驶、人脸识别、智慧医疗以及智能风控等多个场景中的落地应用并且取得了巨大的成功.在许多任务中,当呈现自然发生的输入时,机器学习模型的表现甚至胜过了人类.

然而,大多数的机器学习模型在设计时并未考虑攻击者的存在.尽管在预测正常样本时模型能有优异的表现,但在现实场景中,由于可能存在大量的恶意用户甚至是攻击者,机器学习模型在生命周期的各个阶段都可能面临着不同程度的安全风险,导致模型无法提供正常的服务或者是泄露模型的隐私信息.例如,攻击者可能对模型的训练数据和输入样本进行恶意篡改或是窃取模型参数,从而破坏模型的机密性、可用性和完整性,这就是机器学习模型面临的安全与隐私问题.

为了构建安全可靠的机器学习系统,消除机器学习模型在实际部署应用中的潜在安全风险,保证机器学习模型的机密性、完整性和可用性,一大批来自学术界和工业界的学者系统地研究了机器学习模型安全与隐私问题,并且前瞻性地提出了一系列针对模型安全和隐私的对抗攻击和防御方法,涵盖了机器学习模型的整个生命周期.然而,由于不同学者所处的研究领域不同,解决问题的角度不同,因而构建的威胁模型也不同,所提的攻击或防御方法也各有侧重.因此,我们亟需要对现有的研究工作进行系统的整理和科学的归纳、总结、分析,以便为后续学者了解或研究机器学习模型安全提供指导.

在本文中,我们首先详细地阐述了机器学习中的CIA模型.然后,我们从数据安全、模型安全以及模型隐私三个角度对现有的攻击和防御研究进行系统的总结和科学的归纳,并讨论了相关研究的局限性.最后,我们讨论了机器学习模型安全与隐私研究所面临的挑战以及未来可行的研究方向.

1 机器学习中的CIA模型

随着人工智能安全研究的进一步深入,机器学习模型安全与隐私问题逐渐引起了学术界的关注.Papernot等人将机器学习模型安全需求总结为三个特性:机密性(Confidentiality)、完整性(Integrity)和可用性(Availability),即机器学习中的CIA模型.机器学习模型机密性要求机器学习系统必须保证未得到授权的用户无法接触到系统中的私密信息,即包括模型的训练数据,也包括模型的架构、参数等信息;完整性要求模型的预测结果不能偏离预期;可用性则要求机器学习系统在面对异常输入甚至是恶意输入时仍能够提供正常服务.然而,现有研究表明,在机器学习模型生命周期的各个阶段,机器学习的CIA三个特性都有可能被攻击破坏,所对应的攻击方法分别称为机密性攻击、完整性攻击和可用性攻击.

机密性攻击. 机器学习即服务(Machine Learning as a Service, MLaaS)平台为大量非专业的数据持有者训练模型提供了便利,但这同时也可能会泄漏数据持有者的隐私数据.文献[1]指出,MLaaS平台上由第三方提供的模型未必可信.当数据持有者使用MLaaS平台时,可能会选到由攻击者精心设计的恶意模型.例如,攻击者可以将训练数据编码到模型参数中,然后通过解码参数窃取用户的隐私.此外,文献[2]中提出了一种基于解方程形式窃取模型参数的攻击方法,在此基础上,攻击者可以基于模型逆向的方法生成与模型私密训练数据相似的数据,从而对模型的机密性造成巨大威胁.

可用性攻击.由于模型推理阶段机器学习系统可能会接收并处理大量的异常输入甚至是恶意的输入,因而机器学习模型的可用性也可以成为攻击者的攻击目标,以迫使系统无法提供正常的服务.例如,在无人驾驶领域,如果攻击者把一个非常难以识别的东西放在车辆会经过的路边或者是对交通标志进行物理意义上的扰动,就有可能迫使一辆自动驾驶汽车进入安全保护模式并停在路边,无法进行正常工作.

完整性攻击.完整性攻击发生在模型的训练阶段或预测阶段.在训练阶段,最常见的攻击方式是投毒攻击[3],即攻击者通过篡改训练数据或添加恶意数据来影响模型训练过程,最终降低其在预测阶段的准确性.在预测阶段,最典型的攻击方式是对抗样例攻击,即攻击者通过在测试数据中添加精心构造的微小扰动达到让模型预测出错的目的.

2 数据安全风险与保护

机器学习模型除了预测阶段容易受到对抗样例攻击之外,其训练过程本身也可能遭到攻击者攻击.特别地,如果机器学习模型是根据潜在不可信来源的数据(例如 Yelp、Twitter 等)进行训练的话,攻击者很容易通过将精心制作的样本插入训练集中来操纵训练数据分布,以达到改变模型行为和降低模型性能的目的[4-6].这种类型的攻击被称为“数据投毒”(Data Poisoning)攻击,它不仅在学术界受到广泛关注,在工业界也带来了严重危害.例如微软 Tay,一个旨在与 Twitter 用户交谈的聊天机器人,仅在 16 个小时后被关闭,只因为它在受到投毒攻击后开始提出种族主义相关的评论.这种攻击令我们不得不重新思考机器学习模型的安全性.

2.1 投毒攻击

最早关于投毒攻击的研究可追溯到[7,8].Newsome 等人[8]设计了一种攻击来误导检测恶意软件中的签名生成.Nelson 等人[4]表明,通过在训练阶段学习包含正面词汇的垃圾邮件,可以误训练垃圾邮件过滤器,从而使其在推理阶段将合法电子邮件误分类为垃圾邮件.Rubinstein 等人[9]展示了如何通过注入干扰来毒害在网络传输上训练的异常探测器.Xiao 等人[10]研究了 LASSO、岭回归(Ridge Regression)和弹性网络(Elastic Net)三种特征选择算法对投毒攻击的鲁棒性.在恶意软件检测任务上的结果表明,特征选择方法在受到投毒攻击的情况下可能会受到严重影响,例如毒害少于 5%的训练样本就可以将 LASSO 选择的特征集减弱到几乎等同于随机选择的特征集.

Mei 等人[11]证明了最优投毒攻击可以表述为一个双层优化问题,并且对于某些具有库恩塔克(Karush-Kuhn-Tucker,KKT)条件的机器学习算法(例如支持向量机、逻辑回归和线性回归),利用隐函数的梯度方法可以有效地解决这一问题.Alfeld 等人[12]针对线性自回归模型提出了一个通用的数学框架,用于制定各种目标、成本和约束条件下的投毒攻击策略.Jagielski 等人[5]对线性回归模型的投毒攻击及其防御方法进行了系统研究,并提出了一个特定于线性回归模型设计的理论基础优化框架.除了传统的机器学习模型之外,投毒攻击还被扩展至深度神经网络[13]、强化学习[14]、生物识别系统[15]以及推荐系统等[16][17].Muñoz-González 等人[13]提出了一种基于梯度优化思想的投毒攻击算法,大大降低了攻击复杂度.Suciu 等人[18]提出了 StringRay,这种方法不仅在四种分类任务上成功地实现了定向投毒攻击,同时能够绕过两种现有的防御机制[19,20].

Fig.1 A stop sign and its backdoored versions using, from left to right, a sticker with a yellow square, a bomb and a flower as backdoors

图 1 停车标志及其受后门攻击的版本,后门触发器(从左到右)为黄色方块、炸弹和花朵[21]

最近,备受学界关注的“后门攻击(Backdoor Attack)”[21,22]或“木马攻击(Trojan Attack)”[23]就是一种危害性更大的投毒攻击,它使攻击者能够将“后门”或“木马”植入到模型中,并在预测阶段通过简单的后门触发器完成恶意攻击行为.被植入“后门”的深度神经网络在正常样本上表现很好,但会对具有特定后门触发

器的输入样本做出特定的错误预测.“后门”可以无限期地保持隐藏直到被带有特定后门触发器的样本激活,隐蔽性极强,因而有可能给许多安全相关的应用(例如生物识别认证系统或自动驾驶汽车)带来严重的安全风险[21-23].例如,Gu 等人[21]通过将带有特殊标签(即后门触发器)的“停车”标志图像插入训练集中并标记为“速度限制”以在路标识别模型中生成后门.该模型虽然可以正确地分类正常街道标志,但会对拥有后面触发器的恶意停车标志产生错误的分类.因此,通过执行此攻击,攻击者可以通过在模型上贴上标签来欺骗模型,将任何停车标志归类为速度限制,从而给自动驾驶汽车带来严重的安全隐患.虽然后门攻击和对抗样例攻击都会导致模型误分类,但对抗样例的扰动特定于输入和模型,而后门攻击则可以使攻击者能够选择最方便用于触发错误分类的任何扰动(例如,在停止标志上贴标签).此外,后门攻击也可被用来给深度神经网络加上“水印”,将模型识别为特定供应商的知识产权,以防止具有商业价值的模型被轻易复制[24].

2.2 防御方法

大多数针对投毒攻击的防御机制依赖于一个事实,即投毒样本通常在预期输入分布之外.因此,投毒样本可被视为异常值,并且可以使用数据清理(即攻击检测和删除)[20]和鲁棒学习(即基于对边远训练样本本质上不太敏感的鲁棒统计的学习算法)[5]来净化训练样本.

鲁棒学习. Rubinstein 等人[9]利用稳健统计的知识构建了一个基于主成分分析(Principal Component Analysis, PCA)的投毒攻击检测模型.为了限制异常值对训练分布的影响,该检测模型约束 PCA 算法搜索一个特定方向,该方向的投影最大化了基于鲁棒投影跟踪估计的单变量离散度量,而不是标准偏差.Liu 等人[25]假设特征矩阵可以很好地用低秩矩阵来近似,并在此基础上集成了稳健低秩矩阵近似和稳健主成分回归方法用于稳健回归.受稳健统计中利用修剪损失函数来提高鲁棒性这一做法的启发,Jagielski 等人[5]提出了一种名为 TRIM 的针对回归模型的防御方法,并提供关于其收敛的正式保证以及在实际部署时投毒攻击影响的上限.在每次迭代中,TRIM 使用具有最低残差的子集计算修剪版的损失函数.本质上,这种方法是在对抗环境中应用经过修正的优化技术进行正则化线性回归.

数据清理. Shen 等人[3]针对不能接触到所有训练数据的间接协作学习系统提出了相应的防御方法 Auror,这种方法首先识别与攻击策略对应的相关掩蔽特征(Masked Features),然后基于掩蔽特征的异常分布来检测恶意用户.Steindhardt 等人[26]尝试在训练模型之前检测并剔除异常值来防御投毒攻击,并在经验风险最小化的情况下,得出了任意投毒攻击影响的近似上限.Baracaldo 等人[27]利用 tamper-free provenance 框架[28],提出利用训练集中原始和变换后数据点的上下文信息来识别有毒数据,从而实现在潜在的对抗性环境中在线和定期重新训练机器学习模型.Zhang 等人[29]提出一种利用一小部分可信样本来检测整个训练集中的恶意样本的算法(DUTI).具体地,该方法寻求针对训练集标签的最小更改集,以便从该校正训练集学习的模型能正确地预测可信样本的标签.最后,该方法将标签被更改的样本就被标记为潜在的恶意样本,以提供给领域专家人工审核.

后门攻击检测.模型后门攻击检测极具挑战性,因为只有当存在后门触发器时才会触发恶意行为,而后门触发器在没有进一步分析的情况下通常只有攻击者知道.因此,无论是提供训练数据的用户还是提供预训练模型的用户,都无法保证其基于机器学习模型的相关操作的安全性.为解决这一挑战,Chen 等人[30]提出了激活聚类(Activation Clustering, AC)方法,用于检测被植入后门触发器的训练样本.该方法通过分析训练数据的神经网络激活状态,以确定它是否遭受后门攻击以及哪些数据样本是恶意的.Wang 等人[31]提出了针对深度神经网络后门攻击的检测系统,利用输入过滤、神经元修剪和 unlearning 等方法能够识别深度神经网络中是否存在“后门”并重建可能的后门触发器,从而保证模型在实际部署应用中的安全性.

3 模型安全风险与保护

近年来,机器学习、深度学习等核心技术已被广泛应用于图像分类、语音识别、自动驾驶、垃圾邮件过滤以及智能反欺诈等现实任务.研究表明,攻击者试图通过各种方法改变模型输入特征以绕过现实任务中的机器学习模型的检测,或直接对模型进行攻击以破坏其完整性,从而达到对抗目的.其中,攻击者最常用的攻击手段是通过向正常样例中添加精细设计的、人类无法感知的噪音来构造对抗性样例,从而达到不干扰人类认知而促

使机器学习模型对精心构造的对抗性样例作出错误判断的目的,这种攻击方法被称为“对抗攻击”或者是“对抗样例攻击”.以图像分类为例, 如图 2 所示,原始图片以 57.7%的置信度被图像分类模型识别为“熊猫”,而添加细微扰动之后得到的对抗性图片则以 99.3%的置信度被错误地识别为“长臂猿”,然而对于人而言,对抗性图片依然可以正常地被识别为大熊猫.由于这种细微的扰动通常是人眼难以分辨的,因而使得攻击隐蔽性极强,但其足以改变模型的预测结果,危害性极大,因而给现实场景中尤其是风险敏感场景中实际部署应用的机器学习模型带来了巨大的安全威胁.

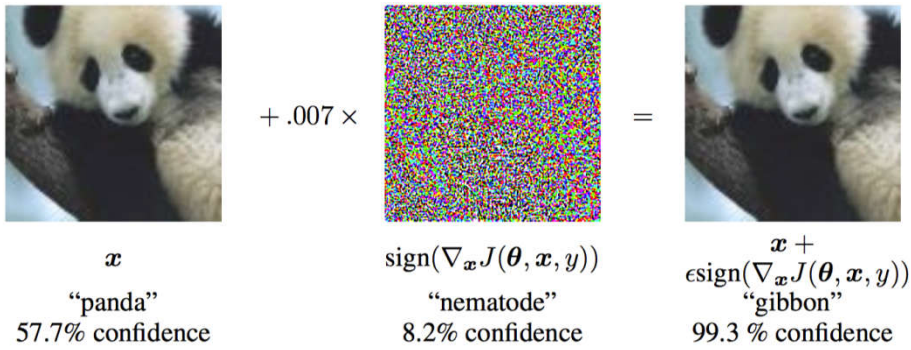

Fig.2 An example of adversarial attack

图 2 对抗样例攻击示例[32]

与其他攻击不同,对抗性攻击的核心在于如何构造能促使机器学习模型产生误分类的对抗样例,因而主要攻击过程发生在对抗样例构造阶段.一旦构造完成,该对抗样例便如同正常样例一般被攻击者输入到目标攻击模型中以误导模型的决策过程,从而达到欺骗待攻击模型的目的.在对抗样例的构造过程中,根据攻击者所获取到的目标模型具体信息的多少,对抗攻击可以分为白盒对抗攻击和黑盒对抗攻击.

- **白盒攻击**.白盒攻击假设攻击者可以完全获取目标模型的结构和参数等信息,因而在攻击过程中,攻击者可以利用模型的完整信息求解目标模型的梯度信息,以指导对抗样例的生成过程.
- **黑盒攻击**.与白盒攻击不同,黑盒攻击假设攻击者既无法得知目标模型采用训练数据和模型结构,也无法获取模型的具体参数,只能获取模型的最终决策结果.在这种情况下,待攻击模型对于攻击者而言犹如一个黑箱,攻击者只能通过操纵模型的输入和利用最终决策结果来探测目标模型的敏感性或对模型的梯度信息进行数值估计,以进而指导对抗样例的构造过程.因而,相较于白盒攻击,黑盒攻击所利用的信息更少,攻击的难度更大.

3.1 对抗样例理论研究

Szegedy 等人[33]在 MNIST 数据集上的实验表明,在测试集上表现优秀的分类模型其实并没有从训练数据中学到符合正确决策结果的内在特征,并且这种现象具有普遍性.虽然这些模型在自然数据上表现优秀,但当测试样本在整个分布中低概率出现时,这些模型就暴露出了缺陷.因此,Szegedy 认为对抗样例存在的原因之一是模型的高度非线性导致的输入与输出映射的不连续性,以及次优的模型平均和次优的正则化导致的过拟合.

然而,Goodfellow 认为对抗样例的存在是高维空间中的线性特质导致的[32].在高维线性模型空间中,输入数据的多个微小变化叠加会导致输出的极大变化,即如果线性模型的输入维度足够高,那么它就容易受到对抗样例攻击.对于深度神经网络等非线性模型,为了保证模型易于训练,通常会选择 ReLU 等分段线性激活函数.即使是采用 Sigmoid 激活函数,通常也会让神经元尽可能处于非饱和的区域.因此,非线性模型中的线性行为也使得模型的完整性易受对抗样例攻击.

最近的一项研究表明[34],对抗样例的产生可归因于非稳健特征的出现:某些来自数据分布模式的特征对于人类来说是难以理解的,但它们具备高度预测性.同时,他们也对对抗样例的迁移性给出了解释:由于任意两

个模型可能同时学习到类似的非稳健特征,因此扰动此类特征的对抗样例可以对二者同时产生影响.

3.2 对抗样例攻击方法

作为破坏机器学习模型完整性最强有力的攻击方法,对抗样例攻击被广泛应用于诸如计算机视觉、自然语言处理、音频处理以及图数据处理等各个领域.

3.2.1 计算机视觉

在计算机视觉领域,对抗攻击旨在通过向图片中添加人眼无法感知的噪音以欺骗诸如图像分类,目标识别以及看图说话等多种机器学习模型.根据在攻击过程中是否依赖模型具体的结构和参数信息,针对计算机视觉模型的对抗攻击方法可以分为白盒攻击和黑盒攻击.为了保证攻击的隐蔽性,无论是白盒攻击还是黑盒攻击均需要限定所添加扰动的幅度,从而保证促使学习模型产生误分类的同时,不干扰人的识别.

3.2.1.1 白盒攻击

基于优化. Szegedy 等人[33]首次提出“对抗样例”这一概念,将寻找最小可能的攻击扰动定义为一个优化问题,并用提出使用 L-BFGS 来解决这个问题.这种方法攻击成功率很高,但同时计算成本也较高.Carlini 等人[36]进一步改进了 L-BFGS 方法,提出了攻击效果更好的目标函数,并通过改变变量解决边界约束问题,这一方法通常被称为 C&W 攻击.Chen 等人[37]在 C&W 攻击的基础上结合弹性网(Elastic Net)正则思路提出了 EAD,该方法生成的对抗样本相较于C&W生成的对抗样本具有更强的迁移性.Khrulkov等人[38]提出了一种基于求解优化问题的构造通用扰动的新算法,该算法主要基于深度神经网络特征映射的雅可比矩阵的(p,q)-奇异向量(Singular Vectors).

基于梯度.为了降低计算成本,Goodfellow等人[32]提出了快速梯度符号法FGSM,这种方法假设在数据点附近决策边界是线性的,因此沿着梯度的反方向添加扰动即可拉大对抗样例与原始样本的距离.这种方法虽然能快速的生成对抗样例,但在实际情况中,由于线性假设往往不成立,这种方法无法很好的拟合模型;此外,FGSM是一种单步(One-step)攻击方法,因此攻击成功率较低.为了进一步提升 FGSM 的攻击效果,Kurakin 等人[39]提出了基本迭代方法 I-FGSM(或 BIM),使用贪婪法在每次迭代中将对抗样本沿梯度方向移动.然而,迭代方法生成的对抗样本很容易过拟合到局部极值点,因此迁移性没有单步攻击生成的对抗样例强[40].为了解决这个问题,Dong 等人[41]提出了基于梯度的动量迭代攻击方法 MI-FGSM,在稳定更新的方向时又能逃离局部极值点,使得生成的对抗样本具有很高的可迁移性,进而使其具有强大的黑盒攻击能力.Xie 等人[42]在 MI-FGSM 的基础上引入了输入转换(Input Diversity)并提出了 M-DI^2-FGSM,进一步提高了对抗样本的迁移性.此外,Madry 等人[43]发现I-FGSM可以通过ε范围球内的随机点开始而得到显著的改善,因此提出了一种名为PGD的攻击方法,有效地提升了 I-FGSM 的攻击效果.Zheng 等人[44]将 PGD 推广至数据分布空间,使学 d 习得到的对抗样例分布能最大程度地增加模型的泛化风险.Papernot 等人[45]提出了基于雅可比矩阵的 JSMA 方法,其主要思想是通过添加稀疏噪音的方式来构造对抗样例.这种方法则是允许添加大的扰动,但是要求被扰动的像素点要尽可能的少.

基于分类超平面.尽管 FGSM 等基于梯度的对抗样例攻击方法能快速地生成是原分类器产生误分类的对抗样本,但这类攻击方法存在一个共性问题,即无法控制达到攻击目标的最优扰动规模.为解决这一问题,Moosavi-Dezfooli 等人[46]提出 Deepfool 算法,目的是寻找可以使分类器产生误判的最小扰动.在此基础上,Moosavi-Dezfooli 等人[47]还提出了一种通用的、不依赖于某一特定样本的对抗扰动(Universal Adversarial Perturbation,UAP)生成方法,能使所有被添加该扰动的所有图片都会被误分类为其他类别.相比于基于梯度信息的对抗样本生成方法,基于分类超平面的方法所生成的扰动具有更强的泛化能力和更强的黑盒攻击能力.

基于生成模型.Baluja 等人[48]提出对抗性转换网络(Adversarial Transformation Network,ATN),它能够将任何输入样本转换为使目标网络产生错误分类的对抗样例,同时对原始输入和目标网络输出的干扰最小.Song 等人[49]提出基于条件生成模型(Conditional Generative Models)的对抗样例生成方法,其主要思想是首先通过

训练辅助分类器生成对抗网络（AC-GAN）以对数据样本的条件分布进行建模,然后以目标类别为条件,在生成器的潜在空间上搜索被目标分类器错误分类的图像.为了生成感知上更真实的对抗样本,Xiao 等人[50]提出一种基于 GAN（Generative Adversarial Network）的对抗样例生成方法 AdvGAN,其中生成器用于产生对抗扰动,鉴别器用于确保生成的对抗样例是真实的.特别地,生成网络一旦训练完毕,就可以有效地为任何样本生成扰动而不再需要查询目标模型.

对抗补丁.Brown 等人[51]放宽了“扰动必须是人眼不可察觉的”这一限制,提出“对抗补丁（Adversarial Patch）”生成算法,使其加到任何图片上都可以让原图被识别为特定类别.Liu 等人[52]提出 PS-GAN,将 GAN 和 Grad-CAM[53]结合到对抗补丁的训练中去,以训练一种更不易被发现但又拥有强攻击力的补丁.Thys[54]等人针对目标检测系统提出了一种对抗补丁生成算法,并且这种对抗补丁能够在真实世界中拥有物理攻击效果.

其他.Xiao 等人[55]首次提出了通过空域变换来生成对抗样本,即通过改变原始样本中像素点的位置来生成对抗样例.虽然该方法在传统的对抗样本生成评价指标中和原图像会有较大的 Lp 距离,但是从人的视觉感官上这种变换方式更真实,且更不容易被现有对抗攻击防御方法检测出来.从这项研究中我们可以得出一个新的结论,即利用 L2 距离作为原始图像与对抗样例的相似性度量不符合人的视觉感受机制.Su 等人[56]提出单像素攻击,即通过只改变一个像素点的值来使模型分类出错.

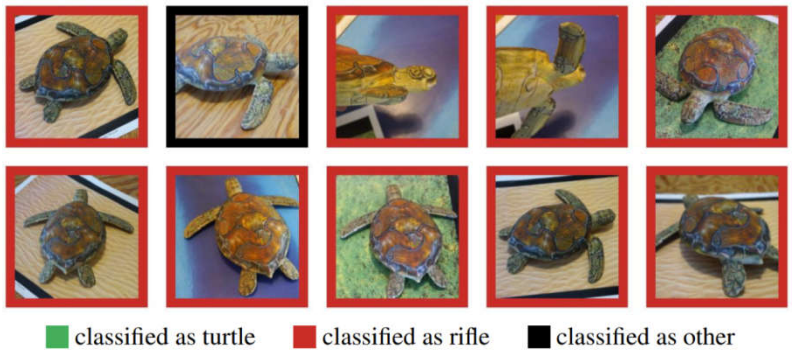

Fig.3 Different random poses of a 3D-printed turtle perturbed by EOT are classified

图 3 模型对 EOT 生成的 3D 打印乌龟的不同随机姿势进行分类[57]

物理世界的实际攻击.大部分上述对抗样本在现实世界的危害有限,因为数据会受变焦、相机噪声、角度和距离等其他的影响.Kurakin 等人[39]首次研究了物理世界的实际攻击方法,并讨论了通过摄像头实际拍摄给对抗样本带来的影响.Athaly 等人[57]对物理环境下的对抗攻击进行了更加深入的研究,探讨了 2D、3D 和物理世界 3D 这三种环境下的对抗样本的生成方法和有效性问题,并首次制作了能在各个角度下欺骗分类模型的真实 3D 物体.该研究提出一种通用的对抗样本生成方法——变换期望算法（Expectation Over Transformation,EOT）,通过在优化过程中对不同干扰进行建模,使得该方法生成的对抗样本在模糊、旋转、缩放、光照等变换下都表现出很强的鲁棒性.Eykholt 等人[58]提出了一种通用的攻击算法 RP_2（Robust Physical Perturbations）,其能够在不同的物理条件下产生鲁棒的对抗扰动.

3.2.1.2 黑盒攻击

由于在模型的实际部署应用中,我们通常无法获取模型的架构、参数等信息,只能操纵模型的输入和输出,因此在这种场景中,黑盒攻击更具有普遍性和现实意义.根据攻击时采用的策略的不同,现有的黑盒攻击方法主要分为基于迁移性的方法[59-63]、基于梯度估计的方法[64-67]、基于决策的攻击方法[68]和基于采样的方法[69].

基于迁移性的方法.相关研究表明对抗样本具有迁移性（Transferability）[59],即针对目标模型生成的对抗样本同样有可能让其他具有不同结构、以不同训练集训练得到的模型出错.因此在黑盒场景下,攻击者可以在与黑盒目标模型相同或具有类似分布的数据集上训练自己的模型,然而针对自己训练的模型生成的对抗样本

并利用其迁移性欺骗黑盒的目标模型.在攻击者无法获取训练数据的情况下,攻击者可以基于模型蒸馏的思想利用目标模型对自己合成的数据打标签,并用合成数据来训练替代模型以近似目标黑盒模型,然后利用白盒攻击方法针对替代模型生成对抗样本,并利用生成的对抗样例对目标模型进行黑盒迁移攻击[61].然而,这种方法虽被证明适用于类内差异性较低的数据集（例如 MNIST）,但尚未有研究证明它可以扩展到 CIFAR 或 ImageNet 等更复杂的数据集.随后,Papernot 等人[60]利用蓄水池算法（Reservoir Sampling）提高了替代模型的训练效率,Ilyas 等人[62]针对查询次数有限、仅给出 top-k 类别概率和仅给出样本类别标签等条件更严格的情况对替代模型攻击方法进行了改进,Shi 等人[63]提出的 Curls&Whey 攻击则从多样性、迁移性、噪声大小等方面进一步优化了基于替代模型的攻击方法.

基于梯度估计的方法.Chen 等人[64]提出基于零阶优化的有限差分算法 ZOO 来直接估计目标深度学习模型的梯度以生成对抗样例.实验结果表明 ZOO 攻击算法显著优于基于替代模型的黑盒攻击算法,并且与白盒算法 C&W 攻击效果相当.然而,这种方法需要较多的查询次数且依赖于模型的预测值（例如类别概率或置信度）,因此无法应用于模型查询次数有限或模型仅给出类别标签的情况.针对模型查询次数有限的情况,Bhagoji 等人[65]利用随机特征分组（Random Feature Grouping）和主成分分析（PCA）算法减少生成对抗样例所需的查询模型的次数,Ilyas 等人[66]将梯度先验（Gradient Priors）与老虎机优化（Bandit Optimization）算法结合来克服这一局限.Tu 等人[67]提出 AutoZOOM 框架,主要包括两个模块：1）为了平衡模型查询次数和失真度的自适应随机梯度估计策略,和 2）用于提升攻击效率的用未标记数据离线训练的自编码器（Autoencoder）或双线性调整操作,该框架应用于 ZOO 攻击算法时可在维持攻击效果不变的情况下大大减少所需模型查询次数.

基于决策的攻击方法.在真实世界的机器学习相关应用中,攻击者很少能够获得模型的预测值.针对目标模型仅给出类别标签的情况,Brendel 等人[68]提出边界攻击（Boundary Attack）算法,其主要思想是将初始化的图像或噪声逐渐向原始样本靠近直到找到决策边界,并在决策边界上找到与原始样本最近的对抗样本.与基于迁移性的攻击相比,它们需要的模型信息更少、实现简单、实用性更强,但是需要巨大的查询次数.在梯度掩蔽、内部随机性或对抗训练等防御方法存在的情况下,这种基于决策的攻击比其他类型的黑盒攻击更难以防御.

基于采样的方法.在 Ilyas 等人[66]提出的攻击方法中,为了使投影梯度法有效,梯度必须对梯度信息进行相对准确地估计.然而,由于部分神经网络的预测函数是不平滑的,因此用自然进化策略（Natural Evolution Strategy, NES）进行梯度估计不够可靠.为了解决这一缺陷,Li 等人[69]使用有约束的 NES 公式作为目标函数,并以正常输入样本为中心的$\ell_p - ball$上定义的概率密度分布来平滑损失函数.如果能够找到一个损失很小的分布,那么从该分布中采样的样本很可能就是对抗样例.该方法不再依赖于梯度估计,因此它不会受到深度神经网络非平滑性的阻碍.

3.2.2 自然语言处理

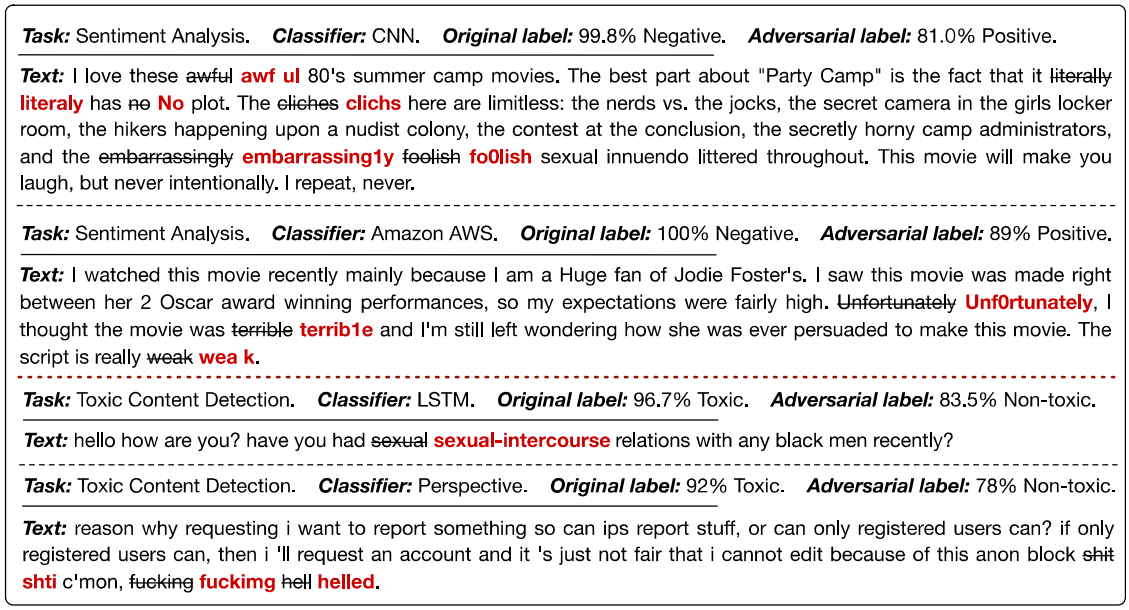

Fig.4 Adversarial examples for text classification

图 4 文本分类的对抗样例[75]

自然语言处理领域的对抗攻击是指在不改变文本语义的情况下使神经网络出现误判.相比于计算机视觉领域,自然语言处理领域的对抗攻击有以下几个难点.首先,由于文本数据是离散的,因此针对图像领域的对抗样例生成方法并不能直接应用于文本.其次,图像的扰动是人眼难以察觉的像素值的微小变化,但是对于文本的对抗攻击,人眼很容易察觉到小的扰动.例如,替换字符或单词会产生无效的单词或语法不正确的句子,并且可能会改变句子的语义.此外,如果直接将图像领域的基于梯度的对抗攻击方法应用到经过向量化处理后的文本特征,生成的对抗样例有可能是无效的字符或单词序列[70].近年来,许多研究者对不同的自然语言处理任务进行对抗攻击,包括问答系统[71]、机器翻译[72]、对话生成[73]、有毒评论检测[74]等.

白盒攻击.Papernot 等人[76]最先开始研究文本序列中对抗样本的问题,提出了一种基于 JSMA 算法思想的对抗文本生成方法,成功攻击了递归神经网络（RNN）.Ebrahimi 等人[77]提出了一种基于梯度优化的白盒对抗文本生成方法 HotFlip,并在随后的工作中将其扩展至定向攻击[78].该方法能够在 one-hot 表示下处理离散文本结构,通过字符替换使字符级文本分类模型出错.Liang 等人[79]基于 FGSM 算法的思想,提出用梯度来度量词语对分类结果的影响程度,并对重要的词语进行插入、删除和修改等扰动.但是,这种方法添加扰动的过程需要人为干预,因此 Samanta 等人[80]将这个扰动过程自动化,并对替换/添加的单词进行限制以使原文的语法结构保持正确.Gong 等人[81]基于 FGSM 和 Deepfool 的思想对词向量（Word Embedding）进行扰动,然后使用词移距离(Word Mover Distance,WMD)找到最近邻词语进行替换.Lei 等人[82]证明了用于文本分类的网络函数的次模性,指出贪婪算法可以很好的近似最优解.

黑盒攻击.Jia 等人[71]首次将对抗攻击应用于问答系统,其具体做法是在段落末尾添加无意义的、分散注意力的句子,这些句子不会改变段落的语义和问题的答案,但会欺骗问答系统.Wang 等人[83]通过改变分散注意力句子的位置来改进了 Jia 等人的工作,并扩展用于生成分散注意力的句子的假答案集.Li 等人[75]提出一种通用的对抗文本生成框架 TextBugger,其核心思想与敏感性分析解释方法类似,具体做法是利用删去某一单词之后模型输出的置信度变化来衡量每个词对分类结果的贡献度,按单词贡献度从高到低采用同义词替换或拼写

错误等方式使模型误分类,同时保证修改后的文本与原文本的语义变化在一定范围内.论文[72,73]提出了更多扰动策略,包括随机交换相邻 token、随机删除停用词、语法错误、反义词等策略.Zhao 等人[70]提出了基于 GAN 的对抗文本生成算法,该算法包括两个关键组件：用于生成伪数据样本的 GAN 和将输入映射到潜在密集空间的逆变器.通过最小化原始输入和对抗性示例之间的重建误差,对原始输入训练这两个分量.但是,这种方法非常耗时.

恶意软件检测.在恶意软件检测领域,对抗攻击被应用于修改恶意软件的特征以规避恶意软件检测模型的检测.例如,研究人员给恶意软件样本中添加一些正常的字符使其看起来更加真实,并不会被系统检测到；攻击者也可以用感染真实 PE 文件、编译含有恶意代码的真实源码、注入二进制代码的方式来绕过检测.Grosse[84]中借鉴 JSMA 方法[45]构造对抗样例,将其从连续可微的空间转移应用到了离散空间中,初步证明了对抗攻击在恶意软件检测领域的可行性.Kreuk 等人[85]修改了 FGSM 的损失函数,使其能更好应用于恶意软件数据的离散性.此外,相关研究者还利用在文件末尾增加字节[86]、插入 API 序列[87]、GAN[88]生成、强化学习[89]的思想生成恶意软件对抗样本.在防御方面,相关研究者利用对抗训练[90]、随机化思想[91]来防御恶意软件对抗样例.

3.2.3 音频处理

不同于自动驾驶等视觉场景,对于现阶段的语音模型来说,非定向攻击并没有太大的威胁性,因为非定向攻击造成的后果并不会威胁用户的隐私、财产或者生命安全.因此,能够对语音系统产生影响、推动其进步的对抗样本必然是以定向为基础的.由于语音识别系统通常需要对输入音频进行大量预处理,因此无法直接将图像领域的攻击方法直接应用于生成对抗音频.

白盒攻击.在先前的研究工作中,Cisse 等人[93]开发了一个通用攻击框架 Houdini,用于攻击包括图像和音频在内的各种模型,但是他们的方法在反向传播求梯度时,无法对音频特征转换进行计算.为了克服这一挑战,Carlini 等人[92]提出了一种白盒场景下基于梯度的定向攻击方法,使得反向传播能够顺利经过特征转换层,开启了学界在定向对抗语音生成方面的探索.该方法通过将给定的任意波形通过添加噪声的方式转换成一段人耳不能区分但会被语音识别系统识别成完全不同的另一段话的新波形,成功攻击了 DeepSpeech 语音识别模型.但是,这种攻击的迁移性非常弱,几乎不能攻击除了目标模型之外的其他语音识别模型.此外,这种攻击方法需要将对抗音频文件直接作为模型输入才有攻击效果,如果用扬声器播放再用麦克收音,则攻击就会完全失效.为了克服这一缺陷,Qin 等人[94]通过利用声学空间模拟器来模拟音频在无线播放时的环境失真,利用听觉掩码（Auditory Masking）的心理声学原理开发出了人耳不可察觉的音频对抗样本,提高了对抗样本在无线播放时的鲁棒性,同时保持任意完整句 100%的针对性成功率.

黑盒攻击.在机器学习模型的实际部署应用中,攻击者通常不知道模型架构或参数,因此研究者进一步对黑盒场景下的对抗语音生成方法进行了研究.Taori 等人[95]提出了一种基于遗传算法和梯度估计的黑盒对抗语音生成方法.Du 等人[96]提出了一种基于粒子群算法的黑盒对抗语音生成方法,成功攻击了语音识别、说话人识别、音频场景识别模型等安全敏感系统.Yuan 等人[97]提出的 Commandersong 成功攻击了科大讯飞语音识别系统,其主要思想将恶意指令的音频特征以一种人耳难以感知的方式嵌入到一段音乐中,使得播放这段音乐时语音识别系统能够识别出恶意指令.但是,这种攻击可以被基于时序依赖关系的防御方法[98]所防御.

3.2.4 图数据处理

针对图数据（Graph Data）的对抗攻击被定义为通过修改给定的图,使图结构或节点属性的变化在限定范围内情况下降低各种图相关算法的性能.针对图数据的对抗攻击在实际应用场景中早有真实案例,比如说在社交网络中,水军通过模仿正常账户进行关注、点赞、评论等行为来降低自己的可疑性,以规避异常检测算法检测从而避免被封号.

Zügner 等人[99]首次对基于属性图（Attribute Graph）的传统模型和图卷积网络（Graph Convolution

Networks)的对抗攻击进行研究,激起了图数据的对抗攻击与防御研究热潮.Dai 等人[100]针对图神经网络(Graph Neural Network)模型,提出了两种对抗攻击方法:在仅给出预测类别的黑盒场景下,作者提出了基于强化学习的黑盒对抗攻击方法;在攻击者可获得模型预测置信度或梯度的白盒场景下,作者提出了基于遗传算法和梯度下降的对抗攻击方法.Chen 等人[101]针对图聚类算法提出两种攻击方法——定向噪声注入(Targeted Noise Injection)和小社区攻击(Small Community Attack).其中,定向噪声注入通过插入边和节点使得原图中的节点与攻击者插入的节点被聚为一类,而小社区攻击主要通过删除节点和边将本应该被聚为一个类的子图拆散成多个类,同时尽可能维持原图中各个节点之间的联系.Bojchevski 等人[102]利用特征值扰动理论的结论,将针对基于随机游走(Random Walks)的网络学习表示(Network Representation Learning)算法的对抗攻击归结为一个双层优化问题.Wang 等人[103]针对协同分类(Collective Classification)这一传统图模型算法,将对抗攻击定义为一个基于图的优化问题以确定需要扰动哪些边.

3.2.5 攻击方法总结

综上所述,经典的对抗样例攻击方法及其满足的属性如表 1 所示.从表中可以看出,目前的对抗攻击方法仍然集中在图像领域,文本、音频和图数据方向的相关研究相对较少,未来可研究的空间较大.在图像领域,大多数攻击者都是在白盒场景下对数字图像做攻击,并且采用 Lp 范数来控制以及衡量噪声大小,以尽可能地减小添加的扰动对人类识别结果的影响程度.然而,未来图像领域的对抗攻击将逐渐从数字领域转变到物理世界,即如何生成能够攻击现实应用的对抗样例.此外,如何提出更好的、更符合人类认知的扰动衡量标准也是一个值得研究的问题.

Table 1 Summary of classic adversarial attacks

表 1 经典的对抗样例攻击方法总结

方法名称	威胁模型		攻击类型		场景		扰动限制	攻击强度	领域
	白盒	黑盒	非定向	定向	数字	物理			
L-BFGS[33]	✓	×	✓	✓	✓	×	ℓ_2	○	图像
FGSM[32]	✓	×	✓	✓	✓	×	ℓ_∞	○	图像
I-FGSM[39]	✓	×	✓	✓	✓	×	ℓ_∞	◐	图像
MI-FGSM[41]	✓	×	✓	✓	✓	×	ℓ_∞	◐	图像
M-DI^2-FGSM[42]	✓	×	✓	✓	✓	×	ℓ_∞	●	图像
PGD[43]	✓	×	✓	✓	✓	×	ℓ_∞	●	图像
JSMA[45]	✓	×	✓	✓	✓	×	ℓ_0	○	图像
DeepFool[46]	✓	×	✓	×	✓	×	ℓ_2, ℓ_∞	◐	图像
UAP[47]	✓	×	✓	×	✓	×	ℓ_2, ℓ_∞	●	图像
C&W[37]	✓	×	✓	✓	✓	×	$\ell_0, \ell_2, \ell_\infty$	●	图像
EAD[38]	✓	×	✓	✓	✓	×	ℓ_1, ℓ_2	●	图像
ATN[48]	✓	×	✓	✓	✓	×	ℓ_2	◐	图像
AC-GAN[49]	✓	✓	✓	×	✓	×	N/A	◐	图像
AdvGAN[50]	✓	✓	✓	✓	✓	×	ℓ_2	●	图像
PS-GAN[52]	✓	×	✓	×	✓	×	ℓ_2	◐	图像
Xiao et al.[55]	✓	×	✓	✓	✓	×	Total Variation	◐	图像
One pixel[56]	✓	✓	✓	✓	✓	×	ℓ_0	○	图像
EOT[57]	✓	×	✓	✓	✓	✓	ℓ_2	●	图像
RP$_2$[58]	✓	×	✓	✓	✓	✓	$\ell_0, \ell_2, \ell_\infty$	●	图像
Papernot et al.[61]	×	✓	✓	✓	✓	×	ℓ_∞	○	图像
Ilyas et al.[66]	×	✓	✓	✓	✓	×	ℓ_∞	?	图像

Curls&Whey[63]	×	✓	✓	×	✓	×	ℓ_2	⊡	图像
ZOO[64]	×	✓	✓	✓	✓	×	ℓ_2	○	图像
AutoZOOM[67]	×	✓	✓	✓	✓	×	ℓ_2	⊡	图像
Boundary[68]	×	✓	✓	✓	✓	×	ℓ_2	●	图像
HotFlip[77]	✓	×	✓	×	✓	×	Cosine Similarity	○	文本
Papernot et al.[76]	✓	×	✓	✓	✓	×	N/A	○	文本
Jia et al.[71]	×	✓	✓	×	✓	×	N/A	○	文本
TextBugger[75]	✓	✓	✓	✓	✓	×	Cosine Similarity	●	文本
Zhao et al.[70]	×	✓	✓	×	✓	×	ℓ_2	⊡	图像/文本
Houdini[93]	✓	✓	✓	✓	✓	×	ℓ_2, ℓ_∞	⊡	图像/音频
Carlini et al.[92]	✓	×	×	✓	✓	×	ℓ_∞	○	音频
Qin et al.[94]	✓	×	×	✓	✓	✓	PSD	●	音频
SirenAttack[96]	✓	✓	×	✓	✓	×	ℓ_∞	⊡	音频
Commandersong[97]	✓	✓	×	✓	✓	✓	ℓ_1	●	音频
Zügner et al.[99]	✓	×	✓	×	✓	×	Specific Budget	⊡	图数据
Dai et al.[100]	✓	✓	✓	✓	✓	×	N/A	●	图数据
Chen et al.[101]	✓	×	✓	×	✓	×	N/A	⊡	图数据
Bojchevski et al.[102]	✓	×	✓	×	✓	×	ℓ_0	⊡	图数据
Wang et al.[103]	✓	×	✓	✓	✓	×	Specific Budget	⊡	图数据

注：✓=满足,×=不满足；扰动限制（用于控制以及度量扰动大小）：$\ell_p(p = 0,1,2, \infty)$=Lp 范数；攻击强度（对抗样例的攻击成功率）：○=低,⊡=中,●=高.

3.3 对抗样例防御方法

传统的模型优化手段如如权重衰减或者 dropout 虽然在一定程度上可以让机器学习模型更加稳健,但通常无法切实防范对抗样本.机器学习模型内在的复杂性使其在预测阶段难以获得对于对抗攻击的鲁棒性,但这种复杂性又是保证模型具有强大的建模能力的必要条件.目前为止,并没有一个能够达到完全令人满意的程度的对抗样本防御方法,因此设计更强的防御方法是未来机器学习模型安全保护研究的重点.

3.3.1 图像预处理与特征变换

由于许多方法产生的对抗性扰动对于人类观察者来说看起来像高频噪声,因此很多研究者建议使用图像预处理作为防御对抗样本攻击的策略,例如 JPEG 压缩（JPEG Compression）[104]、总方差最小化（Total Variance Minimization, TVM）[105]、图像缝合（Image Quilting）[105]、图像深度缩减（Bit-Depth-Reduction）[106]等.Xu 等人[106]提出深度颜色压缩（Depth-color-squeezing）方法来防御对抗样例,其本质思想是对每个像素进行量化.Buckman 等人[107]提出 Thermometer Encoding 防御方法,其本质思想是对每个像素进行离散化,即用二进制向量替换每个像素原来的值.Guo 等人[108]证明利用局部线性嵌入（Locally Linear Embedding,LLE）来对输入数据进行降维能够提高模型的鲁棒性.Prakash 等人[109]基于模型对自然噪声具有鲁棒性这一现象提出像素偏转（Pixel Deflection）防御方法,该方法通过强制使输入图像匹配自然图像统计来抵御对抗性扰动.Akhtar 等人[110]通过训练扰动校正网络（Perturbation Rectifying Network,PRN）来消除对抗扰动,同时利用 PRN 输入输出差值的离散余弦变换来训练检测器,如果检测到扰动就将 PRN 的输出作为模型的输入,反之将原图作为模型的输入.

由于标准去噪器存在误差放大效应（即微小的对抗性噪声可能会被逐步放大而导致错误分类）,为了解决这一问题,Liao 等人[111]提出 HGD 去噪器,该方法的主要思想是将干净图像的 logits 与去噪图像的 logits 之间的差异作为损失函数来训练去噪器.Shen 等人[112]将消除样本的对抗性扰动定义为学习从对抗样本到原始样本的流形映射的问题,在 GAN 框架下利用对抗样本生成与原始样本相似的重构图像,以达到消除扰动的目的.类似地,Samangouei 等人[113]提出 Defense-GAN,其核心思想是利用生成模型来对正常样本的分布进行建模,然后

生成与待预测样本近似的干净样本,并将干净样本送入模型进行预测.Hwang 等人[114]提出了基于 VAE 的净化对抗样例的方法 PuVAE,通过在每个类的流形上投射对抗样例来消除对抗性扰动,并且将最接近的投影作为净化后的样本.Dubey 等人[115]通过对包含数百亿图像的网络图像数据库进行最近邻（Nearest-Neighbor）搜索来对待预测图像进行近似投影,将最近邻图像的预测结果作为待预测图像的结果.

局部攻击是通过仅在特定的局部区域内添加可见对抗性噪声（Localized and Visible Adversarial Noise, LaVAN)而不影响图像中的显著对象的一种对抗攻击.由于这种攻击在特定图像位置引入集中的高频变化,Naseer 等人[116]提出了局部梯度平滑（Local Gradients Smoothing,LGS）方法,具体做法是先估计梯度域中的噪声位置,然后在图片送入深度神经网络之前正则化估计噪声区域的梯度.与其他防御机制相比,LGS 在是迄今为止对 BPDA（Back Pass Differentiable Approximation）防御性能最好的防御方法.

Wu 等人[117]提出一种结合置信度信息和最近邻搜索的框架 HCNN（Highly Confident Near Neighbor）,将低置信度的（即有可能是对抗样例的）样本点嵌入到高置信度区域,以增强模型的鲁棒性.Song 等人[118]发现对于任意攻击类型或目标模型,对抗样例主要存在于训练数据的低概率分布区域.基于这一认知,他们提出了 PixelDefend,通过将对抗样例移回训练数据的高概率分布区域来净化对抗样例.

虽然图像预处理在攻击者不知道防御方法的场景下很有效,但是它在攻击者知道防御方法的场景下几乎无效[119].但是,预处理仍不失为一类吸引人的防御方法,因为该方法可以与其他防御方法协同工作以产生更强的防御效果,且可以在不知道目标模型的情况下减轻对抗样本的危害.

3.3.2 隐藏式安全

隐藏式安全（Security-by-Obscurity）防御机制通过向攻击者隐藏信息来提高机器学习模型的安全性[7,19,120].这种防御方法旨在防御黑盒环境下攻击者通过查询目标模型来改进替代模型或对抗样例的探测机制.典型的防御方法包括:1）增加模型逆向的难度,例如模型融合；2）拒绝攻击者访问有用的梯度信息；3）随机化分类器的输出.

模型融合（Model Ensemble）.He 等人[121]研究发现将现有的多种弱防御策略集成起来并不能作为一种强防御方法,主要原因是自适应的（Adaptive）攻击者可以设计出具有很小扰动的对抗样本来攻破这三种防御方法.Liu 等人[122]结合模型融合与随机化思想提出了 RSE（Random Self-Ensemble）防御方法,其主要思想是在神经网络中加入随机噪声层,并将多个随机噪声的预测结果融合在一起以增强模型的鲁棒性.这种方法相当于在不增加任何内存开销的情况下对无穷多的噪声模型进行集成,并且所提出的基于噪声随机梯度下降的训练过程可以保证模型具有良好的预测能力.然而,如果没有正确组合基分类器,它们可能会降低安全性[123,124].

梯度掩模（Gradient Masking）[61].梯度掩模防御方法试图通过隐藏能够被攻击者利用的梯度信息来进行防御.然而,这种方法并没有提高模型本身的鲁棒性,只是给攻击者寻找模型防御的漏洞时增添了一定的困难,并且已经有研究表明它可以很容易地被替代模型等方法规避[61,119].

随机化（Randomization）.Xie 等人[125]提出在模型前向传播时使用随机化来防御对抗攻击,包括随机调整大小（Random Resizing）和随机填充(Random Padding).尽管最近的研究表明[122]引入随机性可以提高神经网络的鲁棒性,但是 Liu 等人[126]发现盲目地给各个层添加噪声并不是引入随机性的最优方法,并提出在贝叶斯神经网络(Bayesian Neural Network,BNN)框架下对随机性建模,以学习模型的后验分布.Lecuyer 等人[127]提出基于差分隐私的防御方法 PixelDP,其主要思想是在深度神经网络中加入差分隐私噪声层以使网络的计算随机化,以使 l-norm 范围内的扰动对模型输出的分布变化影响在差分隐私保证的范围内.

3.3.3 影响决策边界

Gu 等人[128]借助收缩自编码（Contractive Auto-Encoder,CAE）的思想提出了深度收缩网络（Deep Contractive Network,DCN）,其主要特点是用于训练的损失函数中包含平滑惩罚项（Smoothness Penalty）.虽然

平滑惩罚提高了深度收缩网络的鲁棒性,但同时也会降低其在正常样本上的性能.Szegedy 等人[129]提出名为标签平滑的防御方法,即用软标签替换硬标签来训练模型.这种方法虽然能防御基于 FGSM 方法生成的对抗样例,但不能防御 JSMA 攻击[45].Cao 等人[130]发现对抗样例通常离决策边界很近,并基于这一认知提出基于区域(Region-based)的分类模型,其主要思想是融合以样本点为中心的超立方体中的信息来进行预测.Yan 等人[131]提出 Deep Defense,其核心思想是将一个基于对抗扰动的正则项集成到目标函数中,这个正则项通过使正确分类样本拥有相对较大值、可能错误分类的样本拥有较小值来惩罚对抗性干扰,在不损失模型精度的情况下提高了模型的鲁棒性.Jakubovitz 等人[132]利用神经网络雅可比矩阵的 Frobenius 范数对模型进行正则化作为其常规训练的后处理(Post-processing)步骤,并证明这种方法可以让原始网络在精度变化最小的情况下提高鲁棒性.

对抗训练.对抗训练[33]是最早提出来的一个针对对抗样本的防御方法,该方法将带有正确标签的对抗样本加入原始训练集中共同训练模型以提高模型的鲁棒性.但是,对抗训练容易使模型过拟合于用于产生对抗样例的特定约束区域中去,导致模型的泛化性能下降[40].例如,Moosavi 等人[46]发现如果在训练和攻击时使用不同的方法来生成对抗样例,那么基于对抗训练的模型不再具有鲁棒性.对抗训练的另一个主要缺点是,它倾向于在无意中学习做梯度掩蔽而不是实际移动决策边界,因此仍然容易受到黑盒攻击的威胁.为了克服这一缺陷,Tramèr 等人提出了集成对抗训练[133]的防御方法,即利用多个预训练好的模型来生成对抗样本,然后将这些对抗样本都加到训练集中对模型进行训练.此外,为了将对抗训练应用到大规模数据集上,Kannan 等人[134]提出了一种基于 logit 配对的对抗训练方法,本质上是在传统的对抗训练基础之上加入了一个正则项,最小化对抗样例的 logit 与对应的原始样本的 logit 的差值.虽然后续研究发现对抗训练容易受到盲点攻击(Blind-Spot Attack)[135],但它仍然是目前最有效的防御方法之一[69].

模型压缩.Papernot 等人[35]提出一种基于知识蒸馏(Knowledge Distillation)的防御方法,将大模型压缩成具有更平滑的决策表面的小模型,在提高模型鲁棒性的同时保持预测精度不变.然而,后续研究证明这种防御方法易被攻破[36].Guo 等人[136]证明利用模型修剪(Pruning)来适当提高非线性深度神经网络的稀疏性能提高其鲁棒性,但过度稀疏的模型可能更难以抵抗对抗样例.Zhao 等人[137]发现模型修剪减少了网络的参数密度,对于用原网络作出的攻击有较小防御性,对参数和激活函数的大幅度量化也能使攻击的迁移性变小.

3.3.4 检测型防御

虽然许多机器学习算法基于平稳性假设(即训练和测试数据来自同一分布),但是特征空间中没有训练数据分布的区域可以在训练阶段分配给任何类别而不会显着增加损失,因此这些区域很容易出现对抗样例.基于此,一些相关研究提出利用检测与特征空间中的训练数据分布相距甚远的样本的方法来检测对抗样例[138,139],常见方法包括基于支持向量机[138]、空间一致性信息[140]、图像变换[141]、高斯判别分析[142]、条件生成模型[143]等.

Metzen 等人[144]利用模型的中间层特征训练了一个子网络作为检测器来检测对抗样例,然而相关研究已证明该检测器容易被训练阶段没有遇到过的攻击方法生成的对抗样例所欺骗[145].为了提高检测器的泛化性能,Lu 等人[145]提出了一种更加鲁棒的检测方法 SafetyNet,其主要思想是利用对抗样例和正常样本在深度神经网络特定层的 ReLU 激活函数输出分布的不同来检测对抗样例;Li 等人[139]则提出从卷积神经网络中各层卷积核的输出中提取统计信息,并在此基础上训练了级联分类器区分正常样本和对抗样例.类似地,Zheng 等人[146]发现当深度神经网络对抗样例分类为特定错误类别时,其隐藏层状态与输入相同类别的正常样本所产生的隐藏层状态完全不同.基于这一认知,他们提出了 I-defender 方法,其核心思想是利用深度神经网络隐含层神经元的输出分布作为其内在特征来来检测对抗样例.

Meng 等人[147]提出了一种攻击无关(Attack Agnostic)的防御框架 MagNet,该框架既不需要修改受保护的分类模型,也不需要了解对抗样例的生成过程,因而可以用于保护各种类型的神经网络模型.MagNet 由一个或多个独立的检测器(Detector)网络和一个重整器(Reformer)网络组成:检测器根据深度学习的流形假设(Manifold Hypothesis)来区分原始样本和对抗样本,对于给定的输入样本,如果任何一个检测器认为该样本是

对抗性的,则将其标识为对抗样本并进行丢弃,反之则在将其送入到目标分类器之前利用重整器对其进行重构;重整器则通过重构输入样本以使其尽可能接近正常样本,将对抗样本的流形移向正常样本的流形,从而削弱对抗扰动对目标分类器的影响.Ma 等人[148]利用局部本质维数(Local Intrinsic Dimensionality, LID)来描述对抗样本在对抗子空间中的维度属性,并且证明这些特征可以有效地区分对抗样本.Ghosh 等人[149]提出基于VAE的防御方法,其中VAE的隐向量服从高斯混合先验分布且每个混合分量对应于一个类别,这使得模型能够进行选择性的分类,即将重构误差超过一定阈值的样本视为对抗样例并拒绝对其进行预测.Pang 等人[150]发现当利用 K-density 检测器或其他基于维度的检测器时,用反交叉熵(Reverse Cross-Entropy, RCE)来代替模型训练过程中常见的交叉熵损失函数可以让模型学到更多区分正常样本与对抗样例的特征.

Tao 等人[151]首先利用图像特征与内部神经元的关联性来确定对模型决策起到关键作用的神经元,然后放大这些神经元的影响同时减弱其他神经元的作用以增强模型决策结果的可解释性,最后基于新模型与原始模型的决策结果检测对抗样例.Zhao 等人[152]利用信息几何学的知识对深度学习模型的脆弱性进行了直观的解释,并提出了一种基于矩阵特征值的对抗样例检测方法.具体地,他们计算了深度神经网络带二次型约束的Fisher 信息矩阵,其中最优对抗扰动由第一特征向量给出,脆弱性由特征值反映:特征值越大,模型越容易受到相应特征向量的攻击.为了提升防御方法的泛化性能,Ma 等人[153]分析了深度神经网络模型在各种攻击下的内部结构,并在此基础上提出了利用深度神经网络不变性特征检测对抗样例的方法,该方法能够以超过 90%的准确率和有限的误报率检测 11 种不同的对抗攻击.

3.3.5 鲁棒优化

鲁棒优化的目的是求得一个对于可能出现的所有情况均能满足约束条件的解,并且是的最坏情况下的目标函数的函数值最优.在鲁棒优化中,对抗性的数据扰动可以被视为一种特殊的噪声.Xu 等人[154]表明,至少对于基于内核的类分类器而言,不同的正则化方法相当于假设输入数据上存在不同类型的有界最坏情况噪声.这有效地建立了正规化学习问题和鲁棒优化之间的等价性,从而将计算要求苛刻的安全学习模型(例如,博弈论模型)近似为计算效率更高的模型,即以特定方式正则化目标函数[155,156].最近,研究者还提出了通过模拟相应攻击来正则化梯度的混合方法,以提高深度网络对对抗攻击的安全性[43,157].

Chen 等人[158]提出了一种基于鲁棒优化的算法来提高基于树的模型的鲁棒性,该方法通过在输入特征的最坏扰动情况下对系统的性能进行优化.Raghunathan 等人[159]基于半定松弛(Semidefinite Relaxation)法计算仅包含一个隐藏层的神经网络在最坏情况下的损失上限,并将这个上限与网络参数一起优化.这种方法相当于提供了一个自适应的正则项,可以增强对所有攻击的鲁棒性.Wong 等人[160]提出了一种方法来学习基于 ReLU 的深度神经网络,这些分类器对训练数据上的范数有界对抗扰动具有可证明的鲁棒性.基本思想是考虑范数有界扰动可达到的激活值的凸外部近似(Convex Outer Approximation),然后基于鲁棒优化的思想最小化该外部区域上的最坏情况下的损失.Sinha 等人[161]采用分布式鲁棒优化的原则,通过考虑 Wasserstein ball 中基础数据分布扰动的拉格朗日惩罚公式,利用训练数据的最坏情况扰动来增强模型参数更新,保证了模型在对抗性扰动下的性能.Madry 等人[43]从鲁棒优化的角度研究神经网络的对抗稳定性,利用鞍点公式找到一组神经网络模型的参数,使模型在对抗样例上损失尽可能小,以获得更加鲁棒的神经网络分类器.

3.3.6 基于博弈论

对抗训练[32,33]或提高决策树和随机森林的鲁棒性[162]是一种典型的防御方法.然而,这些防御是启发式的,没有对收敛性和鲁棒性的理论保证.因此,为了克服这些局限性,研究者提出了更为合理的基于博弈论的方法,引入 Nash 和 Stackelberg 博弈进行安全学习,在假设每个玩家都了解对手和博弈的所有情况下,推导出了博弈平衡存在和唯一性的形式条件[163,164].尽管这些方法看起来很有希望,但了解由此产生的攻击策略在多大程度上能够代表实际情况仍是一个悬而未决的问题[165].由于对抗学习不是一个规则明确的博弈,现实世界攻击者的目

标函数可能不符合上述博弈中的假设.因此,有意识地验证真实世界攻击者的行为是否符合假设,并利用所观察到的攻击的反馈来改进攻击策略的定义是一个有趣的研究方向.这些方法的另一个相关问题是它们对大型数据集和高维特征空间的可扩展性,即有效地解决维度灾难问题,因为生成足够数量的攻击样本来正确地表示它们的分布可能会导致计算成本过高.

4 模型隐私风险与保护

机器学习技术的规模化和产业化发展使其已形成一种商业模式,即机器学习即服务（MLaaS）模式.各大互联网公司纷纷推出了商用 MLaaS 平台,为不具备训练能力的普通数据持有者基于持有的数据训练机器学习模型提供了极大的便利.在这种模式下,数据持有者可以利用第三方提供的模型和算法以及平台提供的计算资源,基于持有的数据,训练用于特定任务的机器学习模型,然后对外开放模型调用接口,提供付费预测服务.尽管这种模式给用户训练和发布模型提供了便利,但同时也使得数据持有者的隐私数据面临着泄露的风险.

在这种场景中,攻击者采用的攻击方法为试探性攻击,即通过某种手段窃取模型信息或者通过部分恢复用于训练模型的成员数据的方式来推断用户数据中的某些隐私信息.根据攻击者窃取目标的不同,试探性攻击可以分为训练数据窃取（Training Data Extraction）攻击和模型萃取（Model Extraction）攻击.其中,数据窃取攻击通过获取机器学习模型训练数据的大致分布或根据模型的预测结果推断训练数据中是否包含某个具体的成员数据的方式窃取训练数据中的隐私信息,而模型萃取攻击则通过在黑盒条件下利用特定手段获取目标模型内部构件或者构造一个无限逼近目标模型的替代模型的方式,达到窃取模型信息的目的.这两类攻击分别从数据和模型两个层面破坏了机器学习模型的机密性

4.1 训练数据窃取

由于机器学习模型在训练阶段会不经意地存储训练数据中包含的隐私信息[166],因此攻击者可以通过攻击机器学习模型来获取有关其训练数据的有意义的信息.

数据窃取攻击（Data Extraction Attack）.在遗传药理学研究领域,机器学习技术被广泛应用于分析病人的基因信息和相关治疗记录以辅助医生进行相应的药物治疗.尽管包含病人隐私信息的数据集通常只对研究人员开发,但基于这些数据学习到的辅助诊断模型却往往是公开的,因此亟需保证模型隐私安全以防止泄露病人隐私信息.然而,Fredrikson 等人[168]针对用药推荐系统的研究却发现分析人口统计信息等属性与药物推荐系统输出结果（华法林剂量）之间的相关性,可以逆向推出病患的遗传信息.类似地,Fredrikson 等人[167]发现攻击者利用机器学习模型的预测结果可以重建模型训练时使用的人脸数据,如图 1.具体地,给定一批输入样本 $X=\{x_1,x_2,\dots,x_N\}$,攻击者可以利用 MLaaS 平台提供的模型接口进行查询,得到相应的预测结果 $Y=\{f(x_1),f(x_2),\dots,f(x_N)\}$.攻击者利用$X,Y$训练得到一个与原始模型$f$近似的替代模型$f'$,然后再基于$f'$逆向恢复$f$的训练数据.Song 等人[1]则发现攻击者可以通过在训练阶段将训练数据编码到模型参数中然后在预测阶段对参数进行解码的方式来窃取训练数据.为了解决训练数据敏感性的问题,保护训练数据隐私,Shokri 等人[169]提出了协作式深度学习（Collaborative Deep Learning）模型,其中每个参与者通过本地训练和定期更新、交换参数来构建联合模型,以保护各自训练集的隐私.然而,Hitaj 等人[170]发现任何隐私保护（Privacy Preserving）的协作深度学习其实并没有真正地保护用于训练的人脸数据,其应用于模型共享参数的记录层面（Record-level）上的差分隐私机制对于作者提出的基于 GAN 的攻击是无效的.针对在线学习（Online Learning）场景下的机器学习模型,Salem 等人[171]提出基于生成对抗网络的混合生成网络（BM-GAN）,利用模型在更新前后针对相同样本预测结果的变化来窃取用于更新模型的训练数据信息.

属性推断攻击（Property Inference Attack）.除了窃取具体的训练数据之外,攻击者可以窃取模型训练数据的敏感隐私属性,如用于恶意软件检测模型的训练数据测试环境或某一类数据在训练集中的占比等.Ateniese 等人[172]首次提出了基于元分类器（Meta-classifier）的属性推断攻击并且证明仅提供记录级隐私的差分隐私机制无法有效地防御属性推断攻击.然而,尽管该属性推断攻击方法针对隐马尔可夫模型（HMM）和支持向量机（SVM）有很强的攻击效果,但由于深度神经网络模型的复杂性使得训练元分类器变得困难,导致严重地削

弱了该攻击在深度神经网络上的攻击效果.为了解决这一问题,Ganju 等人[173]提出了一种新的针对全连接神经网络（FCNNs）的属性推断攻击方法,简化了元分类器的训练过程.Melis 等人[174]发现在协同式深度学习模式下,针对训练数据子集的属性推断攻击仍然能够成功.

成员推断攻击（Membership Inference Attack）.成员推断攻击指攻击者利用模型预测结果来推断模型训练数据中是否包含某个训练样本的一种攻击方式,这类攻击方法同样给机器学习安全和隐私带来了严重的威胁.在医疗领域,许多自动医疗诊断系统都是基于病患的隐私信息构建的,一旦这些基于机器学习模型的自动诊断系统遭受到成员推断攻击,必将导致训练数据中包含的病患隐私信息泄露[175].Shokri 等人[176]提出了一种成员推断攻击方法,该方法首先利用训练数据和目标模型返回的预测概率向量及标签训练一个与目标模型架构相似的影子模型（Shadow Model）,以得到某条数据是否属于影子模型训练集的标签；然后将这些数据输入目标模型,利用模型预测接口返回的预测类别、置信度以及该数据是否在训练集中的二值标签训练一个分类模型；最后,给定一条待推断数据,通过将目标模型针对该数据返回的预测概率和标签输入到训练所得分类模型来判断该数据是否属于目标模型的训练数据集.然而,这种攻击基于的假设条件较强（如攻击者必须了解目标模型结构、拥有与目标模型训练数据分布相同的数据集等）,因此攻击实施的成本较高.为解决此问题,Salem 等人[177]放宽了这些关键假设,并且证明改进后的攻击方法能显著地减低攻击成本,因此将给实际部署应用中的机器学习模型的安全和隐私带来更大的威胁.此外,Melis 等人[174]研究发现协同式深度学习系统同样容易遭受到成员推断攻击.

4.2 模型萃取

在 MLaaS 平台上,由于训练数据通常属于商业机密或其中存在敏感信息,因此对外提供付费预测服务的机器学习模型同样具有一定的机密性.然而,由于机器学习模型通常是由一系列的参数决定的,因此通过求解模型参数就可以实现模型萃取.Tramèr 等人[2]发现攻击者理论上只需要通过模型预测接口进行$n+1$次查询就能窃取到输入为n维的线性模型,.类似地,Oh 等人[178]研究表明攻击者可以从一系列的查询结果中逆向提取得到诸如训练数据、模型架构以及优化过程等神经网络的内部信息,而这些暴露的内部信息将有助于攻击者生成针对黑盒模型的更有效的对抗样例,从而显著提高黑盒对抗攻击方法的攻击效果.此外,Wang 等人[179]提出了超参数窃取攻击（Hyperparameter Stealing Attacks）,研究结果证明该攻击适用于诸如岭回归、逻辑回归、支持向量机以及神经网络等各种流行的机器学习算法.

4.3 隐私保护方法

4.3.1 基于差分隐私的数据隐私保护

隐私保护数据分析研究跨越多个学科,历史悠久.随着互联网技术的飞速发展,包含个人信息的电子数据变得越来越丰富,相应的数据收集和管理技术也越来越强大,因此对于一个健壮的、有意义的、数学上严格的隐私定义以及满足这个定义的一类计算丰富的算法的需求也随之增加.而差分隐私（Differential Privacy）则是针对隐私保护数据分析问题量身定制的隐私定义[180],它将隐私定义为添加或移除输入数据中的任何一条记录不会显著影响算法输出结果的一种属性.与隐私保护数据分析目的一致,隐私保护的机器学习要求学习者可以学习到隐私数据集中的数据分布信息,但同时不能过多的泄露数据集中任何一个个体的信息.在这种场景中,为了提供任何形式的有意义的差分隐私,必须随机化机器学习系统的部分管线.这种随机化过程既可以在模型的训练阶段完成,也可以在模型推理阶段通过随机选择模型预测结果来实现.

训练阶段的差分隐私.训练数据随机化的一个典型方法是是数据满足局部差分隐私[181].Erlingsson 等人[182]设计了一种局部差分隐私机制（RAPPOR）,允许浏览器的开发人员在满足隐私前提下收集并使用来自浏览器用户的有意义的统计数据.具体地,RAPPOR 机制在用户将数据发送到用于收集数据以训练模型的集中式服务器时,采用随机响应来保护用户隐私,即用户在响应服务器查询时以 q 的概率返回真实答案或以 1-q 的概率返回随机值.Liu 等人[183]提出了一种保护用户社交网络隐私信息的方法 LinkMirage,该方法通过模糊社交网络

的拓扑结构,从而允许不受信任的外部应用程序能够收集有意义的、具有隐私保护的用户社交网络信息以用于模型训练.其他大多数研究则通过在训练过程中向损失函数[184]、梯度[185]、参数值[169]等添加随机噪声的方式来提供 ε-差分隐私保证.

预测阶段的差分隐私.在模型的预测阶段,可以通过引入随机噪声以随机化模型预测行为的方式,提供差分隐私保证.然而,随着查询数量的增加,引入的噪声量也随之增长,因而导致模型预测的准确性降低.为了克服这一缺陷,Papernot 等人[186]设计了一种保护数据隐私的通用型框架——PATE（Private Aggregation of Teacher Ensembles）,它不仅能够提供正式的差分隐私保障,也提供一定的直观隐私（Intuitive Privacy）保障.具体地,该框架先将训练数据划分成 N 个不相交的子集,然后用这些子集分别训练不同的模型,得到 N 个独立的教师模型,最后在预测阶段通过统计每个教师模型的预测结果并选取票数最高的结果将预测结果聚合起来.如果大部分教师模型都同意某一个预测结果,那么就意味着它不依赖于具体的分散数据集,所以隐私成本很小；但如果有两类预测结果有相近的票数,那么这种不一致或许会泄露隐私信息.因此,作者在统计票数时引入了拉普拉斯噪声,把票数的统计情况打乱,从而保护隐私.事实上,每次查询聚合教师模型时都会增加隐私成本,因为它每次给出的结果或多或少都会透露一些隐私信息.因此,作者利用聚合教师模型以隐私保护的方式对未标记的公共数据进行标注,然后用标记好的数据训练学生模型,最终将学生模型部署到用户设备上.这种做法可以防范攻击者窃取隐私训练数据,因为在最坏情况下攻击者也只能得到学生模型的训练数据,即带有隐私保护标注信息的公开数据.

防御成员推断攻击. Salem 等人[177]认为成员推断攻击之所以能够成功,原因之一在于机器学习模型在训练过程中普遍存在过拟合现象.基于这一认知,作者提出了利用随机失活（Dropout）和模型集成（Model Stacking）的方法来防御成员推断攻击.Nasr 等人[187]引入了一种隐私机制来训练机器学习模型并将其形式化为最小-最大博弈优化问题,利用对抗性训练算法使模型的分类损失和成员关系推理攻击的最大增益最小化,以使攻击者无法区分最终训练所得模型对其训练数据以及对同一分布中其他数据点的预测结果.Hagestedt 等人[188]则提出了一种新的差分隐私机制 SVT^2,能够显著降低 DNA 甲基化（DNA Methylation）等生物医学数据的成员隐私风险.

4.3.2 基于密码学的模型隐私保护

密码学是数学和计算机科学的分支,其原理涉及大量的信息理论,密码学相关技术被广泛的应用于通信加密及信息完整性验证以保证通信信息的机密性和完整性.在机器学习领域,同态加密、安全多方计算等技术也被广泛地应用于保护机器学习模型的安全和隐私.

Dowlin 等人[189]将同态加密技术引入到神经网络中,以允许神经网络在不解密数据的情况下直接处理加密数据.由于同态加密技术将给机器学习模型的体系结构设计引入额外的约束.因此该方法受限于同态加密的性能开销以及所支持的有限算术运算集.为解决这一问题,Liu 等人[190]为神经网络中诸如线性转换、激活函数和池化等常用操作设计了不经意（Oblivious）协议,并结合乱码电路、同态加密等密码学相关理论提出了 MiniONN,这种方法可以在不需要改变模型训练方式的情况下将普通神经网络转换为不经意神经网络（Oblivious Neural Networks）以支持保护隐私的模型预测.

此外,许多学者提出将安全多方计算（Secure Multi-party Computation）应用于协同式机器学习框架中（例如岭回归[191]、线性回归[192]等）,以保证参与各方的训练数据隐私.Bonawitz 等人[193]提出了一种移动应用场景下的数据聚合安全协议,该协议利用安全多方计算的方式计算自各个用户设备的模型参数更新总和,以确保客户端设备的输入仅由服务器进行聚合学习.该协议不仅开销低,而且还可以容忍大量的设备故障,因此是移动应用的理想选择.Mohassel 等人[194]提出了一种基于安全多方计算的、适用于线性回归、逻辑回归和神经网络的模型训练保密协议,该协议大幅度低提升了已有最先进的解决方案效率.

5 研究难点与未来挑战

尽管机器学习模型安全与隐私研究已经取得了一系列瞩目的研究成果,但目前该研究还处于初级阶段,依然存在许多关键问题尚待解决.同时,万物互联时代数据的持续暴增,深度学习、迁移学习、强化学习等新一代机器学习技术进一步发展应用,给机器学习的安全应用和隐私保护带来了新的挑战.在现阶段,机器学习模型安全与隐私研究面临的主要挑战有:在大数据环境下,如何突破海量多元异构数据的可信处理与隐私保护技术;在对抗环境下,如何进一步增强对抗攻防技术的研究;在开放场景下,如何实现机器学习模型风险量化评估.

5.1 数据可信处理与隐私保护

构建可信、可靠以及隐私保护的数据处理技术体系是保障机器学习模型安全的基石,也是模型安全与隐私保护的上游研究.在大数据环境下,数据具有海量、多元、异构等特点,数据收集也存在着数据来源广、质量不可控、隐私保护要求高等难点,因而给数据可信处理与隐私保护研究带来了巨大的挑战.

数据可信处理与隐私保护的第一个挑战是如何有效增强机器学习模型训练数据的质量,以保证数据的可靠性和安全性.由于机器学习模型训练数据采集可能来自不同数据源,导致其正确性和完整性无法保障,同时,异构数据还可能存在冗余、不一致等问题.而现阶段缺乏多维度的数据评价指标,因而无法对数据质量进行有效的综合的评估.此外,在对抗环境下,攻击者可以制造大量的对抗样本进行下毒攻击来干扰模型的训练过程.然而,对抗样本视觉上通常难以感知,并且攻击手段在不断地演化,而现有研究提出的数据增强与清洗技术只能进行粗粒度的数据处理,无法有效地检测出数据污染中的恶意数据.因此,未来研究应着手建立完善的数据质量评估体系,基于多维度的指标对数据质量进行综合评估,并使用重复消除、缺失处理、逻辑错误检测、不一致数据处理等方法对数据质量进行增强.同时,还需要研究辅助数据的动态检测机制,尤其是基于主动学习策略不断更新对抗样本检测算法,同时在检测出对抗样本的基础上,通过样本的重构实现数据的可信处理,以保证机器学习模型训练数据的可用、可靠、可信和安全.

数据可信处理与隐私保护的另一个挑战是如何突破敏感数据隐私化处理技术,以保证训练数据隐私甚至是训练模型机密性.机器学习数据具有高维度特征,不同特征的敏感程度不一样,对于敏感度高的特征需要进行隐私化处理以免在训练或应用过程中被窃取.现有研究大多数基于差分隐私或同态加密等技术,而在基于差分隐私的隐私化处理技术中,数据的可用性和隐私保证程度之间始终存在权衡,基于同态加密的隐私化处理技术同样受限于所支持的有限算术运算集和加密性能.因此,建立和完善数据敏感性分级评估、分级数据脱敏、数据隐私性评估体系是未来数据隐私保护研究发展的一个必然趋势.

5.2 对抗攻防博弈

现有研究中所提出的对抗攻击算法大多数基于很强的假设,即要求攻击者必须能获取模型的结构、参数等信息以用于计算模型的梯度信息,即使无法获取模型的结构和参数信息,攻击者也必须要在能获取到模型的预测概率的前提条件下才能执行相应的攻击.然而,在实际应用中这些假设条件通常很难满足,由于模型不对外公布,攻击者基本上无法获取模型的具体信息,因此执行相应的黑盒攻击.一旦模型在推理阶段只提供预测结果而不提供对应的置信度概率,那么很多的攻击方法必将失效.此外,现有的对抗攻击研究主要集中在视觉、文本以及语音等领域,针对图数据的对抗攻击研究相对较少.在已有的研究中,对于扰动以及扰动约束的定义主要基于传统的图论概念和模型,而缺乏可解释性以及与实际应用之间的联系.因此,如何弱化现有的强攻击假设,以设计出更鲁棒、更实用的攻击方法,同时将现有的攻击方法扩展到如随机游走(Random Walk)、信念传播(Belief Propagation)传统图模型算法以及图神经网络是未来对抗攻击研究中一个比较有前景的方法.

对抗防御研究随着对抗攻击理论和技术的进步而不断深入,促进对抗攻击研究的进一步发展目的在于促进建立更加完善的对抗防御体系.在现阶段,对抗攻击呈现出动态演化的趋势,在对抗攻防博弈中,对抗防御研究明显处于劣势.具体表现在于现有对抗防御研究所提出的防御算法大多数是被动的静态经验性防御,无法有效地适应对抗攻击方法的演化周期.因此,未来对抗防御研究应着手于建立动态自适应的防御体系,结合对抗环境下攻击与防御的动态博弈理论,提出攻防一体的对抗攻击检测与防御机制,以突破对抗攻击检测机制的动态

演化与自适应防御技术,保证非受控环境下机器学习模型的安全性和可靠性.

5.3 模型风险量化评估

在机器学习模型生命周期中,普遍存在训练数据污染、训练过程劫持、中间数据篡改等问题.同时,由于机器学习模型应用场景多元、算法设计复杂、开发人员先验知识差异,导致模型在设计开发过程中可能本身就存在漏洞和缺陷,此外模型实现所依赖的第三方框架(如 TensorFlow,sklearn 等)同样可能存在内存访问越界、空指针引用等多种软件漏洞,从而给现实部署应用的机器学习模型带来诸如拒绝服务攻击、控制流劫持等潜在危害.然而,现阶段仍缺乏一个完善的风险评估体系,导致无法对机器模型所面临的安全风险进行量化评估,因而无法保证已部署到生产环境中的机器学习模型的安全性和可靠性.

对于模型的原生脆弱性,由于模型实现过程中可能存在的漏洞种类多、逻辑复杂,不同漏洞具有不同的风险系数,使得基于人工规则的漏洞挖掘方式效率低下且不能发现新型漏洞.因而,未来研究需要突破基于零先验知识的模型漏洞自动化挖掘与分析等技术,以构建动态可扩展的模型原生脆弱性分析模型.其中,一种直观的方法是将系统安全领域的模糊测试技术迁移到机器学习领域,通过生成对抗网络等生成模型生成高覆盖率的种子以对模型进行自动化测试,从而突破机器学习模型自动化诊断难题.对于模型面临的外部风险,未来研究需要突破场景相关的模型风险量化评级的难题.我们可以结合具体的应用场景,利用现有的攻防技术评估模型在开放环境中抵御外界风险的能力.

6 结束语

随着机器学习研究进一步发展和机器学习技术在实际场景中的广泛应用,机器学习模型安全与隐私成为了一个新生而又有前景的研究领域,吸引了一大批来自于学术界和工业界的学者的广泛兴趣和深入研究,并且取得了许多瞩目的研究成果.然而,到目前为止,机器学习安全与隐私保护研究还处于初级阶段,依然存在许多关键的科学问题尚待解决.为了重新审视机器学习发展和应用中存在的安全威胁,理清现有研究成果的优势与不足,明确未来研究方向,本文从数据、模型、应用三个层面系统地研究了机器学习模型安全与隐私问题,回顾了大量的极其影响力的研究成果并对相关研究进行了科学的分类、总结和分析.同时,本文指出了机器学习模型安全与隐私保护研究当前面临的挑战,探讨了未来可行的研究方向,旨在为推动机器学习模型安全与隐私研究的进一步发展和应用提供指导和参考.

References:

[1] Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 587-601.

[2] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction apis. 25th {USENIX} Security Symposium ({USENIX} Security 16), 2016: 601-618.

[3] Shen S, Tople S, Saxena P. A uror: defending against poisoning attacks in collaborative deep learning systems. Proceedings of the 32nd Annual Conference on Computer Security Applications, 2016: 508-519.

[4] Nelson B, Barreno M, Chi F J, et al. Exploiting Machine Learning to Subvert Your Spam Filter. LEET, 2008, 8: 1-9.

[5] Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. 2018 IEEE Symposium on Security and Privacy (SP), 2018: 19-35.

[6] Nelson B, Biggio B, Laskov P. Understanding the risk factors of learning in adversarial environments. AISec, 2011, 11: 87-92.

[7] Barreno M, Nelson B, Sears R, et al. Can machine learning be secure?. Proceedings of the 2006 ACM Symposium on Information, computer and communications security, 2006: 16-25.

[8] Newsome J, Karp B, Song D. Paragraph: Thwarting signature learning by training maliciously. International Workshop on Recent Advances in Intrusion Detection, 2006: 81-105.

[9] Rubinstein B I, Nelson B, Huang L, et al. Antidote: understanding and defending against poisoning of anomaly detectors.

Proceedings of the 9th ACM SIGCOMM conference on Internet measurement, 2009: 1-14.
[10] Xiao H, Biggio B, Brown G, et al. Is feature selection secure against training data poisoning?. International Conference on Machine Learning, 2015: 1689-1698.
[11] Mei S, Zhu X. Using machine teaching to identify optimal training-set attacks on machine learners. Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[12] Alfeld S, Zhu X, Barford P. Data poisoning attacks against autoregressive models. Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[13] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017: 27-38.
[14] Ma Y, Jun K-S, Li L, et al. Data poisoning attacks in contextual bandits. International Conference on Decision and Game Theory for Security, 2018: 186-204.
[15] Biggio B, Russu P, Didaci L, et al. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. IEEE Signal Processing Magazine, 2015, 32(5): 31-41.
[16] Fang M, Yang G, Gong N Z, et al. Poisoning attacks to graph-based recommender systems. Proceedings of the 34th Annual Computer Security Applications Conference, 2018: 381-392.
[17] Li B, Wang Y, Singh A, et al. Data poisoning attacks on factorization-based collaborative filtering. Advances in neural information processing systems, 2016: 1885-1893.
[18] Suciu O, Marginean R, Kaya Y, et al. When Does Machine Learning {FAIL}? Generalized Transferability for Evasion and Poisoning Attacks. 27th {USENIX} Security Symposium ({USENIX} Security 18), 2018: 1299-1316.
[19] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning. Machine Learning, 2010, 81(2): 121-148.
[20] Cretu G F, Stavrou A, Locasto M E, et al. Casting out demons: Sanitizing training data for anomaly sensors. 2008 IEEE Symposium on Security and Privacy (sp 2008), 2008: 81-95.
[21] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017.
[22] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526, 2017.
[23] Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks, 2017.
[24] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. 27th {USENIX} Security Symposium ({USENIX} Security 18), 2018: 1615-1631.
[25] Liu C, Li B, Vorobeychik Y, et al. Robust linear regression against training data poisoning. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017: 91-102.
[26] Steinhardt J, Koh P W W, Liang P S. Certified defenses for data poisoning attacks. Advances in neural information processing systems, 2017: 3517-3529.
[27] Baracaldo N, Chen B, Ludwig H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017: 103-110.
[28] Aman M N, Chua K C, Sikdar B. Secure data provenance for the internet of things. Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security, 2017: 11-14.
[29] Zhang X, Zhu X, Wright S. Training set debugging using trusted items. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[30] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728, 2018.
[31] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. Neural Cleanse:

Identifying and Mitigating Backdoor Attacks in Neural Networks, 2019: 0.

[32] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[33] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[34] Ford N, Gilmer J, Carlini N, et al. Adversarial Examples Are a Natural Consequence of Test Error in Noise. arXiv preprint arXiv:1901.10513, 2019.

[35] Papernot N, Mcdaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks. 2016 IEEE Symposium on Security and Privacy (SP), 2016: 582-597.

[36] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP), 2017: 39-57.

[37] Chen P-Y, Sharma Y, Zhang H, et al. EAD: elastic-net attacks to deep neural networks via adversarial examples. Thirty-second AAAI conference on artificial intelligence, 2018.

[38] Khrulkov V, Oseledets I. Art of singular vectors and universal adversarial perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8562-8570.

[39] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[40] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236, 2016.

[41] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9185-9193.

[42] Xie C, Zhang Z, Wang J, et al. Improving transferability of adversarial examples with input diversity. arXiv preprint arXiv:1803.06978, 2018.

[43] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[44] Zheng T, Chen C, Ren K. Distributionally adversarial attack. arXiv preprint arXiv:1808.05537, 2018.

[45] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 2016: 372-387.

[46] Moosavi-Dezfooli S-M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2574-2582.

[47] Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1765-1773.

[48] Baluja S, Fischer I. Learning to Attack: Adversarial Transformation Networks. AAAI, 2018.

[49] Song Y, Shu R, Kushman N, et al. Constructing unrestricted adversarial examples with generative models. Advances in Neural Information Processing Systems, 2018: 8312-8323.

[50] Xiao C, Li B, Zhu J-Y, et al. Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610, 2018.

[51] Brown T B, Mané D, Roy A, et al. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017.

[52] Liu A, Liu X, Fan J, et al. Perceptual-Sensitive GAN for Generating Adversarial Patches, 2019.

[53] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 2017: 618-626.

[54] Thys S, Van Ranst W, Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection. arXiv preprint arXiv:1904.08653, 2019.

[55] Xiao C, Zhu J-Y, Li B, et al. Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612, 2018.

[56] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019.

[57] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.

[58] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945, 2017.

[59] Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770, 2016.

[60] Papernot N, Mcdaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277, 2016.

[61] Papernot N, Mcdaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning. Proceedings of the 2017 ACM on Asia conference on computer and communications security, 2017: 506-519.

[62] Ilyas A, Engstrom L, Athalye A, et al. Black-box adversarial attacks with limited queries and information. arXiv preprint arXiv:1804.08598, 2018.

[63] Shi Y, Wang S, Han Y. Curls & Whey: Boosting Black-Box Adversarial Attacks. arXiv preprint arXiv:1904.01160, 2019.

[64] Chen P-Y, Zhang H, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017: 15-26.

[65] Bhagoji A N, He W, Li B, et al. Practical black-box attacks on deep neural networks using efficient query mechanisms. European Conference on Computer Vision, 2018: 158-174.

[66] Ilyas A, Engstrom L, Madry A. Prior convictions: Black-box adversarial attacks with bandits and priors. arXiv preprint arXiv:1807.07978, 2018.

[67] Tu C-C, Ting P, Chen P-Y, et al. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. arXiv preprint arXiv:1805.11770, 2018.

[68] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248, 2017.

[69] Li Y, Li L, Wang L, et al. NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks. arXiv preprint arXiv:1905.00441, 2019.

[70] Zhao Z, Dua D, Singh S. Generating natural adversarial examples. arXiv preprint arXiv:1710.11342, 2017.

[71] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328, 2017.

[72] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv:1711.02173, 2017.

[73] Niu T, Bansal M. Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models. arXiv preprint arXiv:1809.02079, 2018.

[74] Hosseini H, Kannan S, Zhang B, et al. Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138, 2017.

[75] Li J, Ji S, Du T, et al. TEXTBUGGER: Generating Adversarial Text Against Real-world Applications. arXiv preprint arXiv:1812.05271, 2018.

[76] Papernot N, Mcdaniel P, Swami A, et al. Crafting adversarial input sequences for recurrent neural networks. MILCOM 2016-2016 IEEE Military Communications Conference, 2016: 49-54.

[77] Ebrahimi J, Rao A, Lowd D, et al. Hotflip: White-box adversarial examples for text classification. arXiv preprint arXiv:1712.06751, 2017.

[78] Ebrahimi J, Lowd D, Dou D. On Adversarial Examples for Character-Level Neural Machine Translation. arXiv preprint arXiv:1806.09030, 2018.

[79] Liang B, Li H, Su M, et al. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006, 2017.

[80] Samanta S, Mehta S. Towards crafting text adversarial samples. arXiv preprint arXiv:1707.02812, 2017.

[81] Gong Z, Wang W, Li B, et al. Adversarial texts with gradient methods. arXiv preprint arXiv:1801.07175, 2018.

[82] Lei Q, Wu L, Chen P-Y, et al. DISCRETE ADVERSARIAL ATTACKS AND SUBMODULAR OPTIMIZATION WITH APPLICATIONS TO TEXT CLASSIFICATION, 2019.

[83] Wang Y, Bansal M. Robust Machine Comprehension Models via Adversarial Training. arXiv preprint arXiv:1804.06473, 2018.

[84] Grosse K, Papernot N, Manoharan P, et al. Adversarial examples for malware detection. European Symposium on Research in Computer Security, 2017: 62-79.

[85] Kreuk F, Barak A, Aviv-Reuven S, et al. Deceiving end-to-end deep learning malware detectors using adversarial examples. arXiv preprint arXiv:1802.04528, 2018.

[86] Kolosnjaji B, Demontis A, Biggio B, et al. Adversarial malware binaries: Evading deep learning for malware detection in executables. 2018 26th European Signal Processing Conference (EUSIPCO), 2018: 533-537.

[87] Rosenberg I, Shabtai A, Rokach L, et al. Generic black-box end-to-end attack against state of the art API call based malware classifiers. International Symposium on Research in Attacks, Intrusions, and Defenses, 2018: 490-510.

[88] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. arXiv preprint arXiv:1702.05983, 2017.

[89] Anderson H S, Kharkar A, Filar B, et al. Evading machine learning malware detection. Black Hat, 2017.

[90] Liu N, Yang H, Hu X. Adversarial detection with model interpretation. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1803-1811.

[91] Wang Q, Guo W, Zhang K, et al. Adversary resistant deep neural networks with an application to malware detection. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 1145-1153.

[92] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. 2018 IEEE Security and Privacy Workshops (SPW), 2018: 1-7.

[93] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373, 2017.

[94] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. arXiv preprint arXiv:1903.10346, 2019.

[95] Taori R, Kamsetty A, Chu B, et al. Targeted adversarial examples for black box audio systems. arXiv preprint arXiv:1805.07820, 2018.

[96] Du T, Ji S, Li J, et al. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems. arXiv preprint arXiv:1901.07846, 2019.

[97] Yuan X, Chen Y, Zhao Y, et al. Commandersong: A systematic approach for practical adversarial voice recognition. 27th {USENIX} Security Symposium ({USENIX} Security 18), 2018: 49-64.

[98] Yang Z, Li B, Chen P-Y, et al. Characterizing Audio Adversarial Examples Using Temporal Dependency. arXiv preprint arXiv:1809.10875, 2018.

[99] Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 2847-2856.

[100] Dai H, Li H, Tian T, et al. Adversarial attack on graph structured data. arXiv preprint arXiv:1806.02371, 2018.

[101] Chen Y, Nadji Y, Kountouras A, et al. Practical attacks against graph-based clustering. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 1125-1142.

[102] Bojcheski A, Günnemann S. Adversarial attacks on node embeddings. arXiv preprint arXiv:1809.01093, 2018.

[103] Wang B, Gong N Z. Attacking Graph-based Classification via Manipulating the Graph Structure. arXiv preprint arXiv:1903.00553, 2019.

[104] Das N, Shanbhogue M, Chen S-T, et al. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 196-204.

[105] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117, 2017.

[106] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017.

[107] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples, 2018.

[108] Guo W, Wang Q, Zhang K, et al. Defending Against Adversarial Samples Without Security through Obscurity. 2018 IEEE International Conference on Data Mining (ICDM), 2018: 137-146.

[109] Prakash A, Moran N, Garber S, et al. Deflecting adversarial attacks with pixel deflection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 8571-8580.

[110] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3389-3398.

[111] Liao F, Liang M, Dong Y, et al. Defense against adversarial attacks using high-level representation guided denoiser. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1778-1787.

[112] Shen S, Jin G, Gao K, et al. Ape-gan: Adversarial perturbation elimination with gan. arXiv preprint arXiv:1707.05474, 2017.

[113] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605, 2018.

[114] Hwang U, Park J, Jang H, et al. PuVAE: A Variational Autoencoder to Purify Adversarial Examples. arXiv preprint arXiv:1903.00585, 2019.

[115] Dubey A, Van Der Maaten L, Yalniz Z, et al. Defense Against Adversarial Images using Web-Scale Nearest-Neighbor Search. arXiv preprint arXiv:1903.01612, 2019.

[116] Naseer M, Khan S, Porikli F. Local Gradients Smoothing: Defense against localized adversarial attacks. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 2019: 1300-1307.

[117] Wu X, Jang U, Chen J, et al. Reinforcing adversarial robustness using model confidence induced by adversarial training. arXiv preprint arXiv:1711.08001, 2017.

[118] Song Y, Kim T, Nowozin S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. arXiv preprint arXiv:1710.10766, 2017.

[119] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420, 2018.

[120] Huang L, Joseph A D, Nelson B, et al. Adversarial machine learning. Proceedings of the 4th ACM workshop on Security and artificial intelligence, 2011: 43-58.

[121] He W, Wei J, Chen X, et al. Adversarial example defense: Ensembles of weak defenses are not strong. 11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17), 2017.

[122] Liu X, Cheng M, Zhang H, et al. Towards robust neural networks via random self-ensemble. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 369-385.

[123] Corona I, Biggio B, Contini M, et al. Deltaphish: Detecting phishing webpages in compromised websites. European Symposium on Research in Computer Security, 2017: 370-388.

[124] Biggio B, Corona I, He Z-M, et al. One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time. International Workshop on Multiple Classifier Systems, 2015: 168-180.

[125] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991,

2017.

[126] Liu X, Li Y, Wu C, et al. Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network. arXiv preprint arXiv:1810.01279, 2018.

[127] Lecuyer M, Atlidakis V, Geambasu R, et al. Certified robustness to adversarial examples with differential privacy. arXiv preprint arXiv:1802.03471, 2018.

[128] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068, 2014.

[129] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2818-2826.

[130] Cao X, Gong N Z. Mitigating evasion attacks to deep neural networks via region-based classification. Proceedings of the 33rd Annual Computer Security Applications Conference, 2017: 278-287.

[131] Yan Z, Guo Y, Zhang C. Deep Defense: Training DNNs with Improved Adversarial Robustness. Advances in Neural Information Processing Systems, 2018: 419-428.

[132] Jakubovitz D, Giryes R. Improving DNN robustness to adversarial attacks using Jacobian regularization. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 514-529.

[133] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204, 2017.

[134] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing. arXiv preprint arXiv:1803.06373, 2018.

[135] Zhang H, Chen H, Song Z, et al. The limitations of adversarial training and the blind-spot attack. arXiv preprint arXiv:1901.04684, 2019.

[136] Guo Y, Zhang C, Zhang C, et al. Sparse dnns with improved adversarial robustness. Advances in neural information processing systems, 2018: 242-251.

[137] Zhao Y, Shumailov I, Mullins R, et al. To compress or not to compress: Understanding the Interactions between Adversarial Attacks and Neural Network Compression. arXiv preprint arXiv:1810.00208, 2018.

[138] Melis M, Demontis A, Biggio B, et al. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. Proceedings of the IEEE International Conference on Computer Vision, 2017: 751-759.

[139] Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics. Proceedings of the IEEE International Conference on Computer Vision, 2017: 5764-5772.

[140] Xiao C, Deng R, Li B, et al. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 217-234.

[141] Tian S, Yang G, Cai Y. Detecting Adversarial Examples through Image Transformation. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[142] Lee K, Lee K, Lee H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems, 2018: 7167-7177.

[143] Li Y, Bradshaw J, Sharma Y. Are Generative Classifiers More Robust to Adversarial Attacks?. arXiv preprint arXiv:1802.06552, 2018.

[144] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267, 2017.

[145] Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly. Proceedings of the IEEE International Conference on Computer Vision, 2017: 446-454.

[146] Zheng Z, Hong P. Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks. Advances in Neural Information Processing Systems, 2018: 7913-7922.

[147] Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples. Proceedings of the 2017 ACM SIGSAC

Conference on Computer and Communications Security, 2017: 135-147.
[148] Ma X, Li B, Wang Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint arXiv:1801.02613, 2018.
[149] Ghosh P, Losalka A, Black M J. Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. arXiv preprint arXiv:1806.00081, 2018.
[150] Pang T, Du C, Dong Y, et al. Towards robust detection of adversarial examples. Advances in Neural Information Processing Systems, 2018: 4579-4589.
[151] Tao G, Ma S, Liu Y, et al. Attacks meet interpretability: Attribute-steered detection of adversarial samples. Advances in Neural Information Processing Systems, 2018: 7717-7728.
[152] Zhao C, Fletcher P T, Yu M, et al. The Adversarial Attack and Detection under the Fisher Information Metric. arXiv preprint arXiv:1810.03806, 2018.
[153] Ma S, Liu Y, Tao G, et al. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. 26th Annual Network and Distributed System Security Symposium (NDSS), 2019: 24-27.
[154] Xu H, Caramanis C, Mannor S. Robustness and regularization of support vector machines. Journal of Machine Learning Research, 2009, 10(Jul): 1485-1510.
[155] Demontis A, Russu P, Biggio B, et al. On security and sparsity of linear classifiers for adversarial settings. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2016: 322-332.
[156] Russu P, Demontis A, Biggio B, et al. Secure kernel machines against evasion attacks. Proceedings of the 2016 ACM workshop on artificial intelligence and security, 2016: 59-69.
[157] Lyu C, Huang K, Liang H-N. A unified gradient regularization family for adversarial examples. 2015 IEEE International Conference on Data Mining, 2015: 301-309.
[158] Chen H, Zhang H, Boning D, et al. Robust Decision Trees Against Adversarial Examples. arXiv preprint arXiv:1902.10660, 2019.
[159] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018.
[160] Wong E, Kolter J Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv preprint arXiv:1711.00851, 2017.
[161] Sinha A, Namkoong H, Duchi J. Certifiable distributional robustness with principled adversarial training. stat, 2017, 1050: 29.
[162] Kantchelian A, Tygar J, Joseph A. Evasion and hardening of tree ensemble classifiers. International Conference on Machine Learning, 2016: 2387-2396.
[163] Brückner M, Kanzow C, Scheffer T. Static prediction games for adversarial learning problems. Journal of Machine Learning Research, 2012, 13(Sep): 2617-2654.
[164] Liu W, Chawla S. Mining adversarial patterns via regularized loss minimization. Machine learning, 2010, 81(1): 69-83.
[165] Wooldridge M. Does game theory work?. IEEE Intelligent Systems, 2012, 27(6): 76-80.
[166] Carlini N, Liu C, Kos J, et al. The secret sharer: Measuring unintended neural network memorization & extracting secrets. arXiv preprint arXiv:1802.08232, 2018.
[167] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015: 1322-1333.
[168] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. 23rd {USENIX} Security Symposium ({USENIX} Security 14), 2014: 17-32.

[169] Shokri R, Shmatikov V. Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015: 1310-1321.

[170] Hitaj B, Ateniese G, Pérez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 603-618.

[171] Salem A, Bhattacharya A, Backes M, et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. arXiv preprint arXiv:1904.01067, 2019.

[172] Ateniese G, Felici G, Mancini L V, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. arXiv preprint arXiv:1306.4447, 2013.

[173] Ganju K, Wang Q, Yang W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018: 619-633.

[174] Melis L, Song C, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning, 2019.

[175] Backes M, Berrang P, Humbert M, et al. Membership privacy in MicroRNA-based studies. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016: 319-330.

[176] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP), 2017: 3-18.

[177] Salem A, Zhang Y, Humbert M, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint arXiv:1806.01246, 2018.

[178] Oh S J, Augustin M, Schiele B, et al. Towards reverse-engineering black-box neural networks. arXiv preprint arXiv:1711.01768, 2017.

[179] Wang B, Gong N Z. Stealing hyperparameters in machine learning. 2018 IEEE Symposium on Security and Privacy (SP), 2018: 36-52.

[180] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 2014, 9(3–4): 211-407.

[181] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. Advances in neural information processing systems, 2014: 2879-2887.

[182] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, 2014: 1054-1067.

[183] Liu C, Mittal P. LinkMirage: Enabling Privacy-preserving Analytics on Social Relationships. Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS, 2016.

[184] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. Journal of Machine Learning Research, 2011, 12(Mar): 1069-1109.

[185] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016: 308-318.

[186] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755, 2016.

[187] Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018: 634-646.

[188] Hagestedt I, Zhang Y, Humbert M, et al. MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS). Internet Society, 2019.

[189] Gilad-Bachrach R, Dowlin N, Laine K, et al. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. International Conference on Machine Learning, 2016: 201-210.

[190] Liu J, Juuti M, Lu Y, et al. Oblivious neural network predictions via minionn transformations. Proceedings of the 2017 ACM

SIGSAC Conference on Computer and Communications Security, 2017: 619-631.

[191] Nikolaenko V, Weinsberg U, Ioannidis S, et al. Privacy-preserving ridge regression on hundreds of millions of records. 2013 IEEE Symposium on Security and Privacy, 2013: 334-348.

[192] Gascón A, Schoppmann P, Balle B, et al. Privacy-preserving distributed linear regression on high-dimensional data. Proceedings on Privacy Enhancing Technologies, 2017, 2017(4): 345-364.

[193] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 1175-1191.

[194] Mohassel P, Zhang Y. Secureml: A system for scalable privacy-preserving machine learning. 2017 IEEE Symposium on Security and Privacy (SP), 2017: 19-38.