人工智能安全:威胁和对策

胡玉鹏，邝文新，秦铮，李建利，张继良，湖南大学

高岩松，南京理工大学

李文佳，纽约理工学院

李克勤，纽约州立大学

近年来，随着计算硬件和算法的快速发展，人工智能(AI)在图像识别、教育、自动驾驶汽车、金融、医疗诊断等广泛领域显示出对人类的显著优势。然而，从最初的数据收集和准备，到训练、推理和最终部署，基于人工智能的系统在整个过程中通常容易受到各种安全威胁。在基于人工智能的系统中，数据收集和预处理阶段分别容易受到传感器欺骗攻击和缩放攻击，而模型的训练和推理阶段分别容易受到投毒攻击和对抗性攻击。为了解决这些针对人工智能系统的严重安全威胁，本文回顾了人工智能安全问题的挑战和最新研究进展，从而描绘了人工智能安全的总体蓝图。更具体地说，我们首先以基于人工智能的系统的生命周期为指导，介绍每个阶段出现的安全威胁，然后详细总结相应的对策。最后，还将讨论人工智能安全问题的一些未来挑战和机遇。

CCS概念:•计算方法论→人工智能;敌对的学习;•安全与隐私→系统安全;

附加关键词和短语:对抗性示例攻击、人工智能安全、投毒攻击、图像缩放攻击、数据收集相关攻击

国家自然科学基金批准号:61872130、62122023、U20A20202、62002167、61874042;湖南省交通厅科技项目(批准号201928);湖南省杰出青年自然科学基金(批准号2020JJ2010)、湖南省科技创新领军人才工程(批准号2021RC4019)、福建省自然科学基金(批准号2021J01544)、长沙市重点研发项目(批准号2021J01544)。kq1907103;江苏省国家自然科学基金项目;BK20200461。

作者通讯:胡彦、邝伟、秦正志、李坤、张俊(通讯作者)，湖南大学，湖南 410082;邮箱:yphu@hnu.edu.cn、wenxinkuang@hnu.edu.cn、zqin@hnu.edu.cn、lkl@hnu.edu.cn，zhangjiliang@hnu.edu.cn;南京理工大学，江苏南京210094;电子邮件:马岩松。gao@njust.edu.cn;纽约理工学院李文，纽约，NY 10023;电子邮件:wli20@nyit.edu;李锴，纽约州立大学，奥尔巴尼，NY 12246;电子邮件:lik@newpaltz.edu。

0360 - 0300/2021/11 art20 15.00 美元

https://doi.org/10.1145/3487890

ACM参考格式:

胡，邝文新，秦铮，李 kenli，张纪亮，高岩松，李文佳，李克勤。2021。人工智能安全:威胁与对策。*ACM计算机。Surv. 55,1*，第20条(2021年11月)，36页。

https://doi.org/10.1145/3487890

1介绍

1956年夏天，John McCarthy在达特茅斯会议上首次提出了**人工智能(Artificial Intelligence**, AI)，标志着AI学科的诞生[97]。然而，直到2006年，随着Hinton等人[66]引入深度学习概念，由于计算资源的快速增长、更高效算法的出现以及互联网上数据的爆炸式增长，才迎来了新一轮的AI应用浪潮。

到目前为止，AI技术已经彻底改变了我们日常生活的许多方面[24、31、47、78、88、94、100、140、158、172]，它使我们能够重新思考如何整合信息、分析数据，并利用由此产生的见解来改进整体决策过程。为了在AI领域占据领先地位，国家已经制定了重大的AI战略规划。例如，美国白宫于2016年发布了《国家人工智能研发战略计划》[12]，DARPA于2018年9月宣布未来将投资近20亿美元开发下一代AI技术[147]。此外，中国国务院于2017年发布了《新一代人工智能发展规划》[43]。如今，AI的发展水平已经成为一个国家综合国力的重要体现。

然而，AI的发展必然有两面性，其安全性正在成为一个重大问题，特别是在对安全敏感的基础设施中。根据美国知名科技博客Gizmodo的数据，从2000年到2013年，有144人死于涉及机器人辅助外科医生的手术[33]。从统计上看，2014-2017年亚马逊使用的基于人工智能的招聘工具更倾向于招聘男性，这引发了人们对AI公平性的担忧[37]。2018年3月，Uber的自动驾驶汽车事故引发了人们对AI安全的担忧[134]。此外，Menon等人提出的图像识别算法pulse[102]再次引发巨大争议，有人使用脉冲算法恢复了一张模糊的奥巴马图像，结果却被恢复为白人男性[154]。据The Register报道，法国一款基于GPT-3的聊天机器人建议模型患者自杀[119]。在高风险领域，如自动驾驶、医疗保健、金融等，一个非常微小的错误或漏洞最终可能导致数百万或数十亿美元的损失，甚至有时会导致人命损失。

虽然AI系统通常很“聪明”，但它们也很“脆弱”，这意味着它们很容易被愚弄或攻击。Sun等人[142]、Yakura等人[168]和Zhang等人[174]分别用图形、音频和文本数据讨论了与AI相关的攻击和防御。Chakraborty等人[23]和Ozdag等人[113]概述了与AI模型相关的安全威胁。与其他专注于单一类型数据或AI生命周期的特定阶段的AI安全相关评论不同，我们讨论的安全威胁和对策涉及广泛的典型AI应用，如图像分类、语音识别、**自然语言处理(NLP)**和许多其他场景。此外，我们以AI系统生命周期为线索，探索和分析AI生命周期各个阶段可能存在的安全威胁及其防御措施。

在这项工作中讨论的整体框架如图1所示。值得一提的是，MITRE、微软和其他11个组织联合发布了对抗性**机器学习(ML)**威胁矩阵[42]，这是一个旨在帮助安全的att&ck风格框架

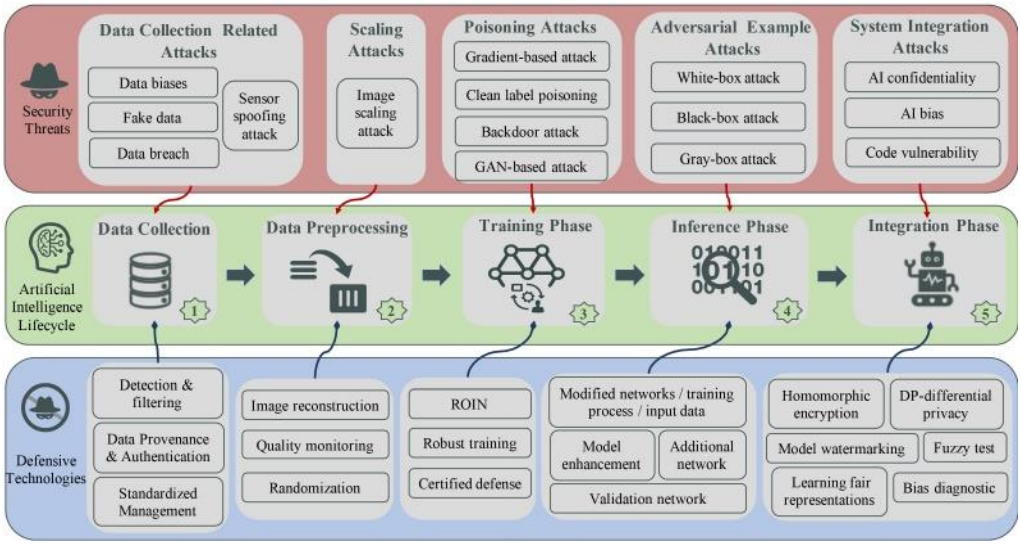

图 1 所示。AI 系统攻击与防御策略的总体框架。

分析师可以快速定位和修复对 ML 系统的攻击。对抗性 ML 威胁矩阵是 ML 系统攻击知识库的第一次尝试，目前正在初步开发中。它包含了特定于 ML 系统的攻击技术，以及适用于 ML 和非 ML 系统的技术。我们的工作可以丰富矩阵，特别是针对 ML 系统的攻击。由于威胁矩阵仍在不断完善，其攻击向量尚未包含最新的攻击技术，如传感器欺骗攻击[132]和**图像缩放攻击**(isa)[**121,161**]。此外，我们的框架还显示了针对不同阶段安全问题的相应对策，可以作为 MITRE 在未来补充防御技术的参考。具体来说，AI 系统的生命周期一般可以分为五个阶段:数据收集、数据预处理、模型训练、模型推理和系统集成，每个阶段都容易受到不同组的安全威胁。

—在数据采集阶段，安全风险与采集数据的方式密切相关。数据收集方式主要有两种:基于软件的收集和基于硬件的收集。基于硬件的收集方法的一种代表性攻击是传感器欺骗攻击[132]，攻击者通过访问或篡改传感器提供的数据来执行传感器攻击[131,132]。基于软件的数据收集方法主要是指收集数字数据，其安全风险包括数据偏差[37,11,154]、虚假数据[30]和数据泄露[38,145]。

-数据预处理阶段。注意目前，缩放攻击一般针对图像域，图像数据在预处理阶段可能被篡改，从而成为潜在的攻击面[76,120,121,161]。具体来说，对于阴险的 isa，攻击者篡改图像并滥用人与机器之间的(视觉)认知差异，以实现欺骗和逃避攻击，甚至绕过仔细的人工检查。与依赖于模型的对抗性示例攻击不同，isa 仅针对数据预处理步骤[121,161]。攻击者利用?$_p$-norm[161]来控制目标图像与攻击图像之间的距离，以提高攻击成功率。数据随机化[161]、质量监控[76]、图像重建[120]是打败 ISA 的主要技术。

-在模型训练阶段，因果攻击通过向模型中注入有毒数据，从而篡改训练模型，从而影响训练数据和训练过程。一般来说，因果攻击主要是指数据中毒攻击[15,27,50,61,93,160]，分为两类，即可用性攻击[15,160]和完整性攻击[61]。对于可用性攻击，通常根据模型的梯度信息[14,77]找到中毒点，或者使用辅助网络自动生成中毒数据[169]。可用性攻击判定-对任何输入的模型的整体性能进行评级。相比之下，完整性攻击不会影响正常输入的分类，而只会影响攻击者选择的输入。后门攻击[62,151]和干净标签中毒攻击[130]是代表性的完整性攻击。现有的防御投毒攻击的策略包括数据消毒[13,77,139]、鲁棒性训练[90]和认证防御[139]。

-在推理阶段，逃避攻击[2,4,18,22,55,136,148,159]通常在模型推理阶段执行，通过制作对抗性示例来降低或干扰模型的预测性能，这些示例通常通过对输入进行较小且语义一致的更改，但不改变目标模型[174]。这种攻击已经在图像分类[18,55,159]、语音识别[2,22,148]、NLP[136]和恶意软件检测[4,84]中得到了广泛的研究。近年来开发了大量的对抗性样例生成策略，如经典的**快速梯度符号法(FGSM)**[55]、**基于雅可比的显著性图攻击(JSMA)[114]**、DeepFool[105]等，主要通过优化搜索或基于梯度的信息实现。相应的，对策也被交互式地设计出来，包括基于模型的策略，如蒸馏[115]、检测器[95]、网络验证[75]，以及基于数据的措施，如对抗性训练[55]、数据随机化[163]和输入重建[36,138]。

-在 AI 系统集成阶段，安全问题变得相当复杂。在实际应用场景中，AI 应用的系统集成不仅涉及到 AI 技术本身的安全风险，还涉及到机载系统、网络、软件、硬件的结合点所产生的问题。这些威胁包括 AI 数据和模型的机密性[48,175]，代码漏洞[162]，AI 偏差[64,154]等。AI的安全需要各领域研究人员的共同努力。

总之，AI的安全威胁已经成为 AI发展和应用中迫切需要解决的问题，特别是对于安全敏感的场景[33,134]。根据攻击所针对的 AI系统的不同阶段，详细阐述了相应的漏洞及其应对措施。

2 .与数据收集相关的攻击和防御

2.1 概述

数据是 AI快速发展的动力。它有许多不同的形式。例如，数据类型包括但不限于:由硬件设备(如传感器)捕获的图像和音频，由计算机系统自动生成的文档和日志，以及由我们的互联网活动产生的那些(如文本、图像、视频、痕迹)。此外，数据收集所涉及的安全问题并不是 AI所独有的，它本质上存在于任何需要数据收集的行业。Lin 等人[89]总结了与网络安全相关的数据收集的要求、目标和技术。他们认为，数据收集需要满足以下安全目标:保密性、完整性、不可否认性、身份验证、隐私保护和自我保护。然而，他们认识到，大多数现有的数据

表 1。数据收集的方法、潜在缺陷和防御措施

方法	安全问题	典型场景	潜在的防御
基于软件的数据收集	数据偏差[37,111,154]	社交网络，推荐系统	检测与过滤[64]，标准化管理
	假数据[30,155]	物联网，社交网络	检测与滤波[16]
	数据泄露[38,145]	涵盖了所需数据收集的场景	加密或认证
基于硬件的数据采集	传感器欺骗攻击[79,131,132,135]	物联网	输入滤波[173]、传感器增强和基带偏移[133]

采集技术满足功能需求，但通常被忽视的安全目标。虽然对数据收集方法的分类还缺乏共识。一般可分为基于软件的数据采集和基于硬件的数据采集[89]。基于软件的数据采集处于数字世界，而基于硬件的数据采集则是将物理世界中的物理量转化为数字形式的关键点。表 1 总结了与数据采集相关的攻击与防御。

*2.1.1 基于软件的数据采集。*互联网用户的日常活动以数字形式产生了大部分数据。数据收集者使用软件程序工具来收集数据(例如，爬虫或“抓取”内容)。基于软件的数据收集需要数据包捕获应用程序、数据包捕获库、操作系统、设备驱动程序和网卡共同工作，以完成数据收集过程。从理论上讲，这个过程中任何一个环节出现问题都会影响数据采集的质量。我们将以在线社交网络为例，讨论基于软件的数据收集方法所带来的安全风险及其相应的防御措施。数据偏差和虚假数据是社交网络数据收集所面临的具有代表性的安全风险。

*2.1.2 基于硬件的数据采集。*与硬件相关的数据采集设备包括传感器、硬件探头、移动终端、数据采集生成卡、内联水龙头、网络接口卡、移动终端等。根据硬件的底层设计原理不同，每种数据采集方式的潜在威胁也不同。传感器是应用最广泛的数据收集工具，它们提供了效率和灵活性的优势。我们以传感器数据采集的安全威胁为例，说明基于硬件的数据采集方法的一些典型安全风险。

2.2 恐怖袭击

*2.2.1 数据偏差。*AI 对训练数据非常敏感。数据源选择和数据准备可能会引入偏差[111]。例如，平台可能受到商业考虑(例如，特定促销)或政治策略的驱动，以“推动”社交网络中的用户行为。此外，社交平台不鼓励第三方收集数据，并对**应用程序编程接口(API)**施加了许多限制。因此，数据收集者只能收集有限的数据或与平台呈现给普通用户的数据不同的数据。

AI 的不完全学习偏见引发了各种各样的担忧，比如性别歧视、种族主义等等。例如，亚马逊人力资源部在 2014 年至 2017 年期间使用了一款支持人工智能的招聘软件[37]。结果，亚马逊公司雇佣了更多的男性求职者，而降低了女性求职者的简历。Twitter 上有人使用 PULSE 算法，将一张输入模糊的奥巴马图像还原为一张白色五官歪斜的新面孔[154]。虽然不是故意的，但 AI 偏见破坏了 AI 的完整性。我们需要改进数据收集标准，开发诊断和减轻偏见的工具。

*2.2.2 假数据。*假数据问题并不是 AI 领域独有的挑战。Wanda 等人[155]创新了卷积神经网络的池化功能。此外，他们提出了一种新的动态**深度神经网络(DNN)**模型算法来检测在线社交网络中的虚假个人资料。Cobb 等人[30]讨论了数据收集应用 **Open data Kit (ODK)**在数据收集过程中的安全挑战。他们探讨了 IDK 数据收集过程中假数据的来源及其防御措施。

*2.2.3 数据泄露。*数据泄露是一个长期存在的问题。Sweeney 等人[145]首先发现，只有三个信息字段(地点、性别、出生日期)可以唯一地识别一半的美国人口。值得注意的是，数据泄露不仅是数据收集阶段特有的问题，也可能发生在模型的训练和推理阶段[38]。

*2.2.4 传感器欺骗攻击。*从物理世界产生的数据需要使用相关的传感器元素进行数字化和收集，用于后续的模型训练和推理。传感器无处不在地集成到智能可穿戴设备、自动驾驶车辆**和光探测与测距(LIDAR)**中，它们是负责数据测量和收集的底层核心组件。攻击者可以利用传感器的物理特性构建恶意样本来欺骗传感器以干扰数据收集[132]。根据目标通道，Shin 等人[131]确定了传感器欺骗攻击的三个向量:常规通道、传输通道和侧通道。Shoukry、Yasser 等人[132]提出了一种通过规则信道进行非侵入式欺骗攻击的方法。为了误导传感器产生恶意速度，攻击者首先阻挡左侧旋转齿轮产生的磁场。检测到错误速度的恶意执行器产生的磁场随后被传输到**防抱死制动传感器(ABS)**，这必然导致传感器欺骗攻击。Foo Kune 等人[79]通过将后门耦合对电路与模拟传感器相结合，进行恶意信号注入，实现了低功率**电磁干扰(EMI)**攻击。音频信号被麦克风拾取。然后输入信号被放大，EMI 通过放大器注入。之后，它们被传输到模数转换器，随后传输到微处理器，最终使电子元件失效。Son 等人[135]利用陀螺仪在自身共振频率下的输出会随噪声波动来攻击**无人机(UAV)**。在陀螺仪的谐振频率下，注入特定噪声会使陀螺仪产生谐振，从而降低精度，干扰 UAV 的操作。

2.3 防 御

数据收集可以从硬件安全、软件安全、网络安全三个方面采取数据安全保护策略，缓解安全威胁。数据收集保护策略种类繁多。此外，保护策略因场景而异。受数据安全策略的启发，我们建议以下三类作为数据收集保护措施。

*2.3.1 检测和过滤。*Hinnefeld 等人[64]研究了 AI 偏差，并设计了一系列策略(例如，优化预处理、拒绝选项分类、学习公平表示和对抗性去权重)来检测和减轻 AI 偏差。为了减轻数据泄露的威胁，Birnbaum 等人[16]提出了一种无监督的离群值检测技术来检测伪造的调查数据，并说明了使用自动数据质量监控的必要性。在硬件数据收集方面，Zhang 等人[173]针对不同的攻击锚点进行了基于软件和硬件的防御。他们发现，传感器增强和基带偏移在防御传感器欺骗攻击方面很有用。以麦克风为例，在放大麦克风幅度的同时增加一个低通滤波器，可以抑制 20 kHz 以上的语音信号，

这意味着人类“听不见”的语音命令将被过滤掉。Ignjatovic 等人[126]证明了聚合多个数据源进行信任评估的传统迭代滤波算法容易受到串通攻击，因此他们提出了一种收敛性更好、鲁棒性更强的迭代滤波技术来保护传感器网络。我们可以丢弃由缺乏可信度和可信度的收集工具捕获的数据。此外，攻击者可以通过记录和重放用户给出的命令来进行欺骗攻击。虽然过滤是一种方便有效的防御手段，但我们需要警惕引入数据偏差的过滤规则。

*2.3.2 数据来源与认证。*适当的传感器信任机制可以适应禁用从不受信任的设备或未经授权的设备收集的数据。首先，在通过可信度评估聚合来自传感器节点的数据之前，应该检查传感器节点的可信度[85,86]。另一种常用的安全机制是身份验证。例如，Shoukry 等人[133]建立了一种物理挑战-响应认证机制 PyCRA，其中传感器使用物理探针连续主动地感知周围环境。认证机制是通过分析主动响应来检测被操纵的模拟信号以防御恶意传感器攻击来实现的。

*2.3.3 标准化管理。*人为的误用也会影响采集数据的质量，这就需要对相关人员进行管理和培训。因此，我们需要审查数据收集的安全要求(保密性、完整性、身份验证等)，并制定相应的管理程序，以确保数据收集的安全性[89,91]。此外，建立适当的激励机制可以鼓励数据提供者更诚实地共享其数据，这有利于数据收集的质量。

3、扩容攻击和防御

3.1 概述

用于训练模型的图像数据的大小通常是固定的。例如，输入模型的图像通常尺寸为 224×224 或 32×32，由于图像预处理步骤的原因，它比原始图像要小。例如，在数据预处理阶段，需要对图像进行缩放以匹配模型输入大小。图像缩放在保留原始视觉特征并按比例缩放的同时，生成比原始图像在像素方面分辨率更低/更高的新图像。然而，在缩放过程中，攻击者可能会滥用缩放算法来调整像素级信息来制作伪装图像，导致图像缩放前后的视觉语义发生巨大变化。如图 2 所示，Xiao 等人[161]基于“羊”图像制作了一幅攻击图像，在视觉上将“狼”伪装成“羊”。一旦对图像进行降采样或调整大小，真实的“狼”就会显露出来。此外，Xiao 等[161]验证了该攻击在多个基于云的图像服务器(如 Microsoft Azure、阿里云、腾讯和百度图像分类服务)上的有效性。例如，百度云服务器以高置信度将图像识别为“狼”。值得注意的是，只要不同的模型使用相同的重新缩放函数来拟合相同的模型输入大小，ISA 就可以对不同的模型不可知。表 2 给出了数据预处理阶段 isa 和防御的概述。

3.2 恐怖袭击

Xiao 等人[161]首先利用插值算法的逆揭示了 ISA。如图 3 所示，首先将扰动矩阵 $_1\Delta$ is 添加到原始图像“srcimg”(例如，数字 8)中，以生成嵌入目标图像(例如，数字 6)的攻击图像“攻击”。而 Δ_2 is 将目标图像“targetimg”与输出图像的差值

图 2 所示。缩放攻击示例[161]。

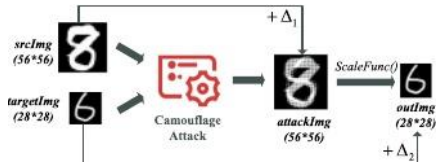

图 3 所示。自动攻击图像制作[161]。

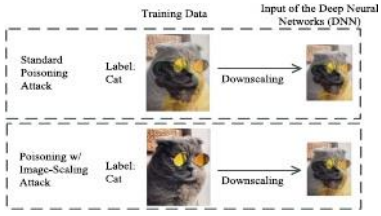

图 4 所示。清洁标签投毒攻击[121]。

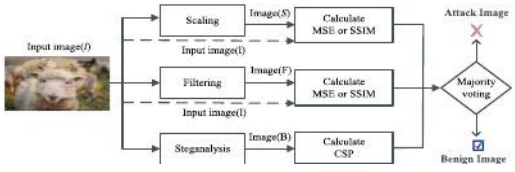

图 5 所示。Decamouflage 概述[76]。

“outimg。”最后，基于 Δ and Δ 约束 1 条件下的插值算法生成最优攻击图像 2。一旦对攻击图像常规执行图像缩放操作，模型就会看到目标图像，例如，位数为 6，从而将其识别为攻击者的目标 6，这是源到目标攻击[161]。即使部署的系统对攻击者来说是一个黑盒，这样的攻击仍然是有效的，因为相对容易推断出所需的参数，如输入图像大小或/和该模型使用的底层重新缩放函数，例如，穷举试验。这是由于通常使用的输入图像大小或/和重缩放函数类型是有限的。

ISA 的根本原因来自于下采样和卷积的相互作用，Quiring 等人[120]从信号处理的角度对其进行了理论分析。他们在三个 ML 成像库(OpenCV、TensorFlow 和 Pillow)上进行了实验，以证实这种相互作用的存在。Quiring 等[121]通过使攻击图像在颜色直方图上与缩放后的图像保持一致，引入了一种新的自适应 ISA，通过检查颜色直方图来降低 ISA 检测的成功率。

此外，作者将 ISA 与投毒攻击相结合，成功隐藏了后门攻击的触发器[121]。如图 4 所示，在数据预处理之前，使用 ISA 技术隐藏触发器，确保带有触发器的攻击图像的内容和标签在视觉上是一致的，从而绕过了对有毒图像的人工检查。作为大多数情况下的标准步骤，一旦执行降尺度操作，触发器就会立即暴露。这种通过图像缩放的毒化样本，保证了触发器的隐蔽性，达到了后门攻击的效果。

3.3 防御

Quiring 等人[120]开发了一种图像重建方法来防御 isa。在他们的工作中使用了选择性中值滤波器和随机滤波器来识别缩放过程中被改变的像素点。然后使用图像中剩余的像素来重建修改过的内容，例如，通过中值。Quiring 等人[120]提出的防御方法避免了对原有神经网络进行修改，而是简单地与已有的图像库结合起来防御缩放攻击。

图像重建通过下采样频率和图像缩放之间的关系来防止 isa，但会降低输入图像的质量[120]。因此，如

表 2。图像缩放攻击和防御

攻击	攻击策略	潜在的防御
肖[161]	减少目标与攻击图像之间的?p-范数距离	随机化, 质量监控
查询[120]	减少目标和攻击图像之间的颜色直方图差距	图像重建[120], 攻击检测[76]

图 5,Kim 等人[76]将缩放、滤波和隐写分析集成到缩放攻击检测框架- decamouflage 中。具体来说，(i)缩放检测方法首先对输入图像进行降尺度操作，然后再进行上尺度操作，构建“复制”图像，然后比较输入图像与其“复制”图像前后在颜色直方图上的相似度:攻击者在输入图像中注入的像素由于上尺度而有望从“复制”中移除。(ii)滤波检测使用滤波器对图像进行滤波。(iii)通过**离散傅里叶变换(DFT)**将疑似攻击图像的样本变换到二维空间，并使用隐写分析检测 ISA 嵌入的扰动像素。随后，使用**均方误差(MSE)**、**结构相似指数(SSIM)**和**中心频谱点(CSP)**指标来量化前后的相似性，并独立推导出每种检测方法的检测边界。最后，执行一种集成技术来识别传入图像是否是攻击图像。

一般来说，查询等[120]消除了攻击效应，但没有具体检测输入图像是否为攻击图像。然而，Kim 等人[76]检测到恶意攻击的存在并拒绝攻击图像。在需要跟踪任何攻击的情况下，检测是首选方法。此外，可以调整[120]，一旦检测到输入为对手，通过重建攻击图像来获得正确的预测，从而进一步消除攻击效应，这可以缓解[120]中图像重建导致的质量下降。

4 数据中毒攻击和防御

4.1 概述

人工智能系统是基于大型策划数据进行训练的。然而，数据质量直接影响训练模型的性能。在这种情况下，攻击者可以毒害训练集来操纵模型的推理行为。从模型和攻击目标的角度来看，投毒攻击可以分为两类:可用性攻击[15,160]和完整性攻击[61]。

- 可用性攻击被称为拒绝服务攻击，其攻击目标是最大化模型的整体损失，并导致模型性能下降以及错误分类。例如，社交媒体聊天机器人拥有丰富的语料库，并通过与人类的交互进行扩展。当攻击者用一些没有上下文相关性的语句影响聊天机器人时，聊天机器人就不会进行正常的逻辑聊天。

- 完整性攻击是攻击者在不影响模型对干净样本的分类的情况下，通过精心设计有毒数据来实现目标损害的攻击[27,61,93]。最具代表性的完整性攻击是后门攻击。后门攻击只会对包含特定(显式甚至不显式)触发器的输入进行错误分类，并且后门仍然可以保留在下游迁移学习任务中。举个后门攻击的例子，在恶意软件检测中，攻击者将包含特定字符串的文件标记为良性数据，并将其放入检测器的训练中。在模型训练和部署后，攻击者只需将特定字符串添加到恶意软件中以逃避检测，因为任何具有特定字符串作为触发器功能的恶意软件都会与良性类相关联。

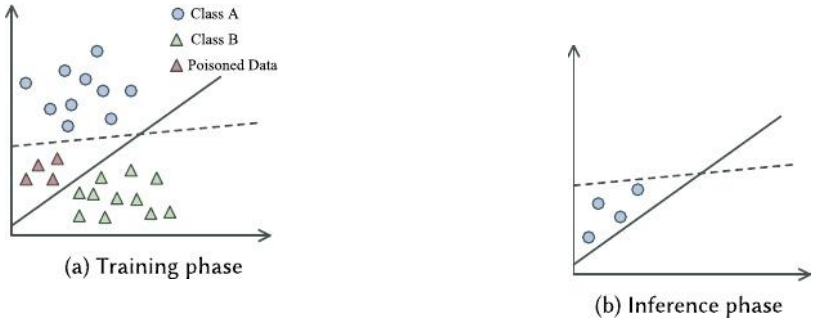

图 6 所示。中毒攻击前后分类模型的变化。

根据攻击行为和分类结果的不同，将投毒攻击分为特定错误攻击和泛型错误攻击。假设有一个干净样本 **C**，真实标签 y_{true}，攻击者构造了一个中毒样本集 C ?并将其添加到模型 M 的训练集中，导致*模型* M 对 *C* 进行错误分类，即 M(C) ?y.*如果* M(C)是攻击者*所* true针对的特定*类*，*则*为特定错误中毒攻击。然而，如果 M(C)是 y 以外的任何类，则它是泛型错误*中毒*攻击。true如图 6(a)所示，实线表示正常情况下的二元分类器。假设在训练集中加入了少量的有毒数据。在这种情况下，决策边界将被移动，从而产生被虚线分隔的分类效果。因此，在正常模型和中毒模型相交形成的封闭区域内的实例将在推理阶段被错误分类。如图 6(b)所示，A 类实例将被错误分类为 b 类。

4.2 恐怖袭击

接下来，我们将详细介绍各种投毒攻击方法，我们也在表 3 中进行了总结。

*4.2.1 可用性攻击。*可用性攻击被称为拒绝服务攻击。代表性的可用性攻击包括基于梯度的攻击和基于**生成式对抗网络 (GAN)**的攻击。带有中毒的可用性攻击可以形式化地表示为双级优化问题[14,15,99]。内部优化是一个在中毒训练集上的模型训练问题。外部优化是最大化攻击者的目标 **A**，通常是由内部优化得到的有毒模型上的干净数据集的分类损失函数 **L**。其形式化表示如下:

吗?吗?吗?吗?吗?吗?

$$Dc* \in \text{argmax}A\ \ Dc?,\ \ \theta\ =\text{l}\ \ d\,?Val,w?,\text{s. t. w}?\in \text{argminl}\ \ ?\text{Dtr}\cup \text{Dc}?,\ \ w_{\text{。}}\ (1)_k D\ \ ?\in \Phi(D)\ w_c$$

攻击者只能访问一个代理数据集 *D*?的数据源。*D* ?分为两个不相交的子集 $D?_{tr}$ 和 $D?_{val}$。$D?_{tr}$ 和中毒样本集 D_c ?，用来训练模型，得到中毒模型参数 *w*。?$D?_{val}$ 用于通过简单的损失函数 $L(D?_{val}, w)$。换句话说，有毒样品对干净数据的影响是由参数 *w* 决定的。?

*Gradient-based 攻击。*基于梯度的投毒攻击的主要挑战是攻击目标相对于投毒点的梯度$\nabla_{xc} A$ *的计算。一般来说，基于梯度的上升和反向梯度优化都是*通过计算攻击目标相对于每个毒化点的梯度来获得优化*后的毒化点 x*。

-基于梯度的上升。梯度上升中毒攻击技术由?? ?

采用梯度上升方法[14,77]。假设攻击函数 $A\ \ Dc?, \theta$

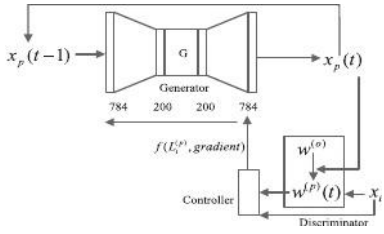

图 7 所示。氮化镓中毒方法综述[169]。

图 8 所示。一个洁净标签中毒的例子。

对于参数 w 和输入 x 是可微的，使用链式法则计算所需的梯度如下:

$$\nabla_x A = \nabla_x L + \frac{\partial w^T}{\partial x} \nabla_w L, \tag{2}$$

其中 $\partial w \partial x$ 表示分类器参数对中毒数据的隐藏依赖。Mei 等人[99]提出了一个隐式方程，该方程使用 **Karush-Kuhn-Tucker(KKT)**条件而不是内部优化问题来推导梯度。通过在中毒点处求导，可以求解梯度。然后，将一个两层优化问题转化为一个单层约束优化问题。虽然优化得到了简化，但梯度计算的复杂性使得它只适用于有限数量的学习算法。

-反向梯度优化。Muñoz-González 等人的反向梯度优化[109]是针对深度学习框架的第一次投毒攻击。他们通过反转学习过程来更新参数。内部优化问题被学习迭代所取代。外部优化问题中所需的梯度是通过内部优化问题中的不完全参数获得的。他们假设一次只优化一个中毒点 $_c$。在内部优化问题中，总共执行 T 次迭代。由此，得到参数 W_T is。利用链式法则计算更新中毒点的梯度。

*基于氮化镓攻击。*Yang 等人[169]设计了一个受 GAN 启发的生成器，以加速有毒样品的产生。生成器首先从干净的训练集 d_i 中随机选择一个样本 x_t 来产生一个有毒样本。然后，鉴别器使用生成器生成的有毒数据来计算干净数据的损失。随后，生成器使用鉴别器提供的梯度和损失的新加权函数更新有毒数据。该过程不断迭代，直到达到终止条件。图 7 概述了基于氮化镓的中毒方法。目标模型作为判别器，而生成器是一个额外的模型，用于生成中毒数据 x_p。从(t− 1)更新 $_{th}$ 中获得的中毒数据被输入到生成器中，以获得在 t 迭代*中* $_p$ 更新的中毒数据 x(*t*)。然后，将 x(t)注入 $_p$ 鉴别*器*，计算加权梯度 f (L, gradient)来更新目标*模型*$^{(p)}{}_i$。每次迭代只需要对目标模型进行一次更新，可以大大减少中毒数据生成的时间。

*4.2.2 完整性攻击。*完整性攻击可以在不影响模型对正常样本分类的情况下完成目标伤害。后门攻击是最具代表性的完整性攻击。

后门攻击。后门攻击[62]并不影响在后门模型中被分类的干净数据的结果，但对于包含由攻击者秘密控制的特定触发器的输入，会产生与预期结果的偏差。借壳攻击是一种典型的完整性中毒攻击，通过在干净样本中添加触发器来创建有毒样本，这些样本的标签通常被修改为目标标签。值得注意的是，触发器，比如它的位置、形状或颜色，都可以在攻击者的任意控制之下。对于输入 x，通过冲压触发器获得其中毒对应的 $A(x,m,\ \Delta)$。以图像域为例，m 表示触发位置。Δ 表示触发颜色、图案等信息。最终的触发器优化问题将两个模型的潜在表示在特征空间中的不相似性最小化，可以将其正式描述为

$$\Delta^f = \underset{\Delta}{\operatorname{argmin}} \sum_{x \in X} \sum_{x_t \in X_t} D\left(F_\theta\left(A(x, m, \Delta), F_\theta\left(x_t\right)\right)\right), \tag{3}$$

*其中，为*目标攻击类，xt 为目标攻击类的*输入*。F(x)表示 *。x 在参数* θ 下的输出或神经网络中某*一层*的*中间*输出。由于经过训练的触发器将激活模型中的一些特定神经元，因此使用 D(•)来测量干净输入和添加触发器的输入在*神经元*激活状态上的差异。常用的 D(•)是 MSE。Gu 等人[61]提出了一种针对神经网络的后门攻击，其中每个神经元都可以看作是一个内部特征。所选神经元所在的层与输出层之间的层被重新训练，使触发器与输出层中的目标神经元类建立强连接。

Turner 等人[151]考虑了一种后门攻击，在这种攻击中，注入的有毒样本在视觉上与标签一致。为了保持标签的一致性，他们通过后门触发幅度修改了原始后门图案的像素值，使后门触发图案在视觉上不明显。实验表明，这种方法可以生成一个不显眼的触发器，并被模型学习，从而实现成功的后门攻击。Barni 等人[9]通过破坏目标类样本数据来添加后门。一旦遇到后门信号，网络就会将样本识别为目标类。该方法允许根据不同的分类任务和目标类选择适当的扰动。例如，对于 MINIST 数字分类任务，他们将基于斜率信号的后门加性扰动定义为 $v(i, j) = j\Delta_m$, $1 \leqslant j \leqslant m$, $1 \leqslant i \leqslant ?M$ 和 l 分别是图像的列数和行数。这种方法形成的后门更加隐蔽，攻击成功率也更高。但是，它们只会破坏目标职业样本，需要提高数据中毒率才能达到较高的攻击成功率。

清洁标签中毒。Shafahi 等人[130]提出了清洁标签中毒攻击，它保留了图像的标签和可视化内容之间的一致性。简而言之，他们通过向训练集中添加有毒数据(标记为基类)来改变模型决策边界，从而导致有毒数据周围的干净目标实例被错误地分类为基类。攻击过程如图 8 所示。首先确定目标类和基类。然后分别从目标和基类中选择一个目标实例 t 和一个基实例 b。在 $?_2$ norm 约束下，以 x 在视觉上与基类相似，但在特征空间表示上接近目标类的方式构造一个中毒样本 x。通过特征碰撞生成的中毒数据表述为

$$p = \underset{x}{\operatorname{argmin}} \|f(x) - f(t)\|_2^2 + \beta \|x - b\|_2^2, \tag{4}$$

其中 $f(x)$是模型倒数第二层 x 的表示，称为 x -特征空间表示的*特征*空间表示。F(x)− F (t)是 $_2{}^2$? 范 $_2$ 数测量

表 3。数据投毒攻击方法

类别	攻击策略	优势	缺点
可用性的攻击	基于梯度的[14,77,109]	模型鲁棒性增强	计算复杂度高
	GAN-based [169]	时间开销低，复杂度低，模型无关	可怜的泛化
完整性的攻击	后门[9,62,151]	不影响正常样本分类、模型无关、高度隐蔽性、高泛化性	局限于图像识别的场景，需要重新训练或引入额外的模型
	清洁标签中毒[130]	简单的实现	计算复杂度高

表 4。数据中毒防御方法

防御策略	优势	缺点
数据消毒[77,110]	通用性强，易于实现，在某些场景下检测成功率高	小样本容易过拟合，计算开销高
鲁棒训练[71,90]	模型鲁棒性增强，低复杂度	高计算开销
认证防御[139]	高可解释性	高复杂性

与目标实例的特征空间相似度，β 调节有毒样本 x 与原始输入空间中基类的视觉相似度。优化问题通过一个正向向后分裂的迭代过程来解决。准确地说，第一步(向前)在特征空间中最小化目标实例和中毒实例之间的$?_2$ 距离。第二步(反向)是最小化输入空间中中毒数据与基础实例之间的距离。遵循式(4)的优化可以提供一组看起来像基类但在深度特征空间中与目标类一致的有毒图像，这样基类标签就不需要改变。

4.3 防 御

数据中毒攻击将有毒数据注入训练集，以破坏学习算法的功能。中毒数据与干净数据具有不同的特征，这意味着中毒数据可以被视为异常，从而可以将异常检测作为防御手段。数据消毒[13,34,77,110,139]通常应用异常检测或模型鲁棒性训练[17,28,71,90,127]可以用来防御数据中毒攻击。我们在表 4 中总结了这些防御措施。

*4.3.1 数据消毒。*Nelson 等人[110]提出了针对垃圾邮件过滤器的数据中毒攻击的**拒绝负面影响(ROIN)**。如果数据对分类器有显著的负面影响，则将其视为中毒数据并从训练集中删除。虽然 ROIN 在某些情况下在防御数据中毒攻击方面表现出色，例如以 100%的成功率识别攻击电子邮件，但在训练集中测试每个数据样本的成本太高。而且，当数据集小于特征数量时，容易出现过拟合。Koh 和 Liang[77]应用鲁棒统计中的影响函数来计算数据点对分类器预测的影响。Koh 和 Liang 提出的方法能够在不重新训练模型的情况下确定每个数据项的影响- ROIN[110]需要重新训练模型-仅使用梯度和 Hession 矩阵，这确保了可以快速识别损害性能的数据点。

*4.3.2 鲁棒性训练。*鲁棒训练通常强烈依赖于一些特征假设。Liu 等人[90]放宽了这些假设，通过改进鲁棒低秩矩阵近似和鲁棒主成分回归，实现了较强的防御性能。Jagielski 等人[71]通过使用修剪损失函数设计了一种名为“TRIM”的对抗性防御技术

在每次迭代中计算残差的不同子集，用于线性回归模型的鲁棒训练。一般来说，**支持向量机**(svm)**对**离群值不具有鲁棒性。Xu 等人[164]改进了相关熵诱导损失函数，并构造了重标度的铰链损失函数来扩展 SVM 的鲁棒性。

4.3.3 认证防御(*Certified 防御*)。Steinhardt 提出了针对投毒攻击的认证防御[139]。为采用异常排除和经验风险最小化的防御者设计了一个框架，旨在研究给定防御的整个攻击空间。假设 $D_{p\ c}$ *and* d 分别表示干净数据集和中毒数据集，θ 表示分类器的参数。相应的防御是为可行集依赖于有毒数据或不依赖于有毒数据 $_p$ 的场景而设计的。以独立于 Das 的情况*为* $_p$ 例，他们提出了一种固定防御方法。在迭代求解过程中，当前最坏攻击点(x,y)= $\arg\max_{(x,y)\in F}{}^{(t)}$?(θ ;x,y)*每次*都*先*$^{(t-1)}$找到。然后在该攻击点的方向上更新模型，得到 θ 。$^{(t)}$最终，就可以找到最坏的中毒*攻击* $_p$ 数据集 d。基于 D 找到最坏验证*错误*上界 M，将其近似为整个*数据* $_p$ *集*(干净和中毒数据)上的*训练*错误。干净数据集上的异常不会过度影响模型。

4.3.4 其他防御措施。为了减轻后门攻击(一种特定类型的数据中毒攻击)的影响，Wang 等人[156]提出了神经净化(Neural cleanup)，其原理是触发最终是将所有图像篡改为目标类所需的(异常)最小扰动。因此，神经净化识别这样最小的扰动，以逆向工程的触发，这可以用来取消后门的移除。Liu 等人[92]在使用干净数据对模型进行微调的同时，修剪了对分类不敏感的神经网络中的冗余神经元，从而使其能够正确分类。然而，他们的方法假设所有模型都可能植入后门，盲目地对模型进行剪枝微调往往会降低正常模型执行正常任务的准确性。Chen 等人[25]基于中毒数据和原始数据在神经网络中神经元激活状态的差异，通过激活聚类技术检测了中毒数据。Gao 等人[51]在不使用任何 ML 技术的情况下，提出了在运行期间检测触发输入的 STRIP。其原理是，对于输入不可知的后门攻击，无论输入内容如何，触发输入将始终被分类到目标标签中。这是因为触发器完全劫持了模型。因此，当在触发器输入中加入强扰动时，预测受到的影响较小:对扰动不敏感。然而，正常输入应该对强扰动敏感。因此，检查一组输入的 perburbed 副本的预测随机性可以区分触发输入和正常输入:触发输入表现出低随机性，而正常输入表现出高随机性。

5 对抗性示例攻击与防御

5.1 概述

在推理阶段进行的对抗性示例攻击是 AI 系统中研究最多的安全威胁。假设有一个经过训练的模型 *m* 预测的 $f(x)=y_{true}$ 的干净样本 x，攻击者*向* x 添加一个难以察觉的扰动 δ 来*创建*一个对抗性示例 x ?= x + δ，*这*足以*误导模型*产生错误的输出 f (x) ?Y，也称为无目标攻击 $_{true}$。在针对性*攻击方面*，就是误导模型产生一个攻击者的目标类 f (x) = y_{target} ?5 .由于 *dnn* 缺乏可解释性和复杂性，AI易受对抗性攻击的原因仍然*没有* $_{true}$ 统一。一些研究人员认为，对抗性攻击的主要原因是

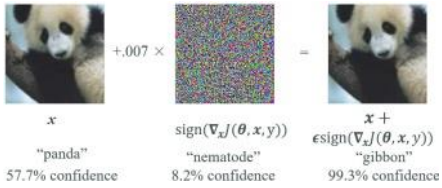

图 9 所示。FGSM 对抗性示例生成概述[55]。

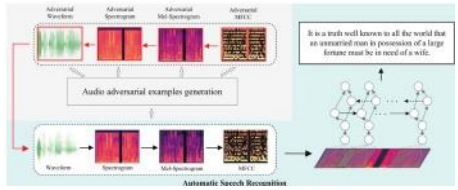

图 10 所示。对抗性攻击和 ASR 过程概述[67]。

模型[116]，而 Goodfellow 等人[55]认为它是由模型在高维空间中的线性行为引起的。Ilyas 等人[70]最近的一项研究表明，对抗性示例与非鲁棒特征密切相关，非鲁棒特征具有高度预测性，但很脆弱，(因此)对人类来说是不可理解的。对抗性示例攻击的问题已经在下面的各个人工智能相关领域进行了演示。

*5.1.1 图像分类。*在图像域，目标是在误导模型的同时，在肉眼无法识别的原始像素上稍微添加扰动[18,35,45,55,80]。对抗性样本的概念最早是由 Szegedy 等人提出的[146]。他们发现，高级神经网络中包含的语义信息(某一特征)分布在整个网络的空间结构中，而不是单个神经元中，神经网络输入和输出之间的映射大多是不连续的。因此，在相同的输入中加入相同量级的干扰，可以使不同的神经网络产生类似的错误分类。具体来说，神经网络有一定的盲点，可以在输入图像中注入有规律的扰动来欺骗网络。Goodfellow 等人[55]提出的**快速梯度符号法(Fast Gradient Sign Method, FGSM)**是图像领域早期具有代表性的对抗样例生成算法。如图 9 所示，对于原始输入 x，所建立的模型将其识别为熊猫 y，在加入精心制作的人类难以察觉的噪声后，网络输出的长臂猿 y 虽然是肉眼可见的熊猫图像，但与原始类不匹配的概率为 99.3%。由于代表图像的像素值近似连续，因此可以人工直观地识别出假图像与合法图像之间的相似性。在其他领域，如语音识别、NLP 和恶意软件检测，数据和分类器结构更复杂，这需要在安装对抗性示例攻击时更加谨慎。

*5.1.2 语音识别。***自动语音识别**(Automatic Speech Recognition,**ASR**)是一种使智能设备能够识别和理解人类语音或/并将其转换为文本的技术，需要在提取声学特征之前对原始音频进行滤波和数字化操作[112]。除了 MFCC 之外，DFT 和**快速傅里叶变换(FFT)**也可以提取语音特征。ASR 的框架如图 10 所示。首先，通过将语音信号分成几个块，这些块可以重叠。然后，通过 FFT 计算每个块的频谱幅值得到频谱图，对频谱进行 mel - filter 得到 Mel-Spectrum。接下来，通过倒谱分析推导出 MFCC[108]。最后，将得到的 MFCC 输入神经网络进行识别/分类。Mel- Filters 用于对频率进行变形，以遵循人耳毛细胞的空间分布。端到端 ASR 系统[3,32,123]通常使用连接**时间分类(Connectionist Temporal Classification, CTC)**损失函数[57]来直接推导字符而不是音素序列。人类和机器语音识别之间的差距产生了不受监控的通道，通过这些通道，对抗性示例可以被植入命令。与图像分类相比较

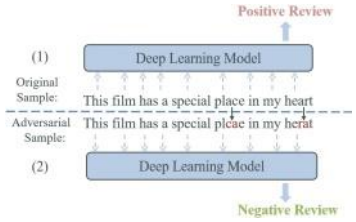

图 11 所示。文本处理示意图中的离散摄动[49]。

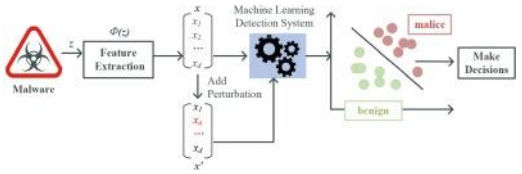

图 12 所示。恶意软件检测对抗示例。

领域，ASR 系统的对抗性示例更难制作，因为一些常见的音频处理操作非常容易引入额外的噪声。

*5.1.3 自然语言处理。*NLP 是用计算机识别人类语言。NLP 的应用范围从输入识别、标签分类到单词、句子的分析理解和处理[54]以及文档的章节[73,96]。对抗性示例攻击也存在于 NLP 中，尽管在这方面的研究少于图像和音频领域。由于图像像素和文本数据之间的差异，像素级对抗性攻击方法不能直接安装在 NLP 中生成。首先，图像数据(例如，像素值)在数值轮廓中是连续的，但文本数据标签类型是离散的。一般来说，文本数据在输入到 DNN 之前需要进行矢量化。词嵌入通常被用作 dnn 的输入。然而，网络可能无法匹配词嵌入空间[54]。其次，对图像的扰动是难以感知的像素值级别的修改。而对于细微的文本扰动，则极易被检测到。如图 11 所示，对原始输入样本(1)中单词的字符进行轻微修改，就会导致单词无效，从而在很大程度上影响句子的整体语义。因此，对于文本类型的对抗性示例攻击，大多数研究关注的是阅读理解任务，而不是短文本。

*5.1.4 恶意软件检测。*恶意软件检测是利用 AI 技术对静态或动态分析提取的软件特征进行分类的过程。静态分析在不执行的情况下提取和分析恶意软件样本的特征。而动态分析则需要执行并分析其对应的特征。常用的动态分析工具包括沙盒、模拟器等。[170]。恶意软件检测中常用的特征有字节序列、操作码、api 和系统调用、网络活动、文件系统、PE 文件等[152]。如图 12 所示，首先，通过特征提取对代表恶意软件的特征序列进行过滤。然后，在由特征序列组成的数据集上训练恶意软件分类器。对抗性示例攻击旨在向特征向量中添加一些功能独立的特征，以生成恶意软件对抗性示例。

5.2 恐怖袭击

攻击者的攻击能力取决于攻击者对模型的了解程度，包括训练数据、特征集、学习算法等。根据攻击者现有的知识，攻击可以分为三大类:白盒、黑盒和灰盒。

-白盒攻击:攻击者对目标模型有充分的了解，包括神经网络模型的类型、参数和训练算法等。攻击者运用已知的知识来识别易受攻击的特征空间，以方便生成对抗性示例。因为白盒攻击需要计算梯度

表 5 所示。对抗性示例生成方法

方法	场景	优势和劣势	攻击特异性	类别
L-BFGS [146]	图像分类	基于优化的搜索，计算复杂度高	目标/没有针对性	白盒
FGSM [55]	图像分类语音识别恶意软件检测	基于梯度，单次迭代，计算简单，扰动大	目标/没有针对性	白盒
I-fgsm /BIM [80]	图像分类	迭代 FGSM，效率更高	目标/没有针对性	白盒
ILCM [80]	图像分类	使最小类的概率增加，迭代求解，本质上类似于 BIM	有针对性的	白盒
JSMA [114]	图像分类自然语言处理恶意软件检测	?0 摄动大的攻击，适合黑盒迁移攻击，需要微观目标模型，难以真实	目标没有针对性	白盒
深愚[105]	图像分类	基于决策面，微扰小，复杂度低;黑箱攻击成功率低	目标/没有针对性	白盒
英国大东电报局[21]	图像分类自然语言处理恶意软件检测	梯度优化求解，具有较高的复杂度	目标/没有针对性	白盒
UAP [104]	图像分类	不需要解决优化问题，梯度计算	没有针对性	白盒
MalGAN [69]	恶意软件检测	Gradient-independent;攻击成功率高，开销高	----	黑盒
EvadeML [167]	恶意软件检测	梯度无关，攻击成功率高，不易扩展	----	黑盒
atn [124]	图像分类	可以攻击一个或多个网络;更高的训练成本	目标/没有针对性	白盒/黑盒
动物园[26]	图像分类	近似梯度估计;成功率高，查询和估计梯度的成本更高	目标/没有针对性	黑盒
胡迪尼[29]	图像分类	欺骗梯度	目标/没有针对性	黑盒
一个像素[141]	图像分类	单像素摄动，无梯度信息，启发式求解，效率低	目标/没有针对性	黑盒

对于输入，虽然梯度在文本情况下是离散的，但基于梯度优化的白盒攻击方法很难应用于 NLP。

-黑盒攻击:与白盒攻击相反，攻击者不知道模型的任何知识，但允许通过查询人工智能系统并仔细设计输入和观察输出来分析模型的漏洞/弱点。黑盒攻击更实用，但设计起来更具挑战性。

-灰盒攻击:灰盒攻击场景由 Meng 等人[101]提出。也被称为半白盒攻击[165]。攻击者需要获得模型的部分知识(除了模型参数)才能完成对目标模型的攻击[142]。灰盒攻击在实践中并不常见[174]。

虽然为了说明当前典型的对抗性示例构建方法，本工作确实涵盖了所有领域的对抗性示例攻击，但在实践中被广泛使用的图像分类是主要的重点，如表 5 所示。除此之外，其他领域的一些攻击将在 5.2.3 节中进行简要讨论。

5.2.1 白盒攻击。

L-BFGS。Szegedy 等人[146]首先证明了人类对图像无法察觉的少量扰动可能会误导神经网络。然而，求解的复杂性

由于对最小扰动的优化问题要求过高，所以他们选择了 L-BFGS 算法来求最小扰动的近似解。具体地说，构造图像 x ?类似于 x:通过将其转化为带有?$_2$-范数约束的凸优化问题*最小化*∇ 22，形式化表示为 $minc\ x- x$?22 + $J_{\#}$ (x?，l) s.t. x ?∈ [0,1]^n 。超参数 $c > 0$。对手试图通过基于?$_p$ -norm 的正则化函数约束对抗性样本与正常样本之间的相似性，同时使用损失函数 $J_{\#}$ (x?，t)使 x ?被误分类为目标类 t。扰动后的对抗性样本仍在正常图像获取范围 x ?∈ [0,1]n。同时，他们证明了将对抗性示例引入

训练过程可以提高模型的泛化能力。

FGSM。Goodfellow 等人[55]设计了 FGSM 来计算对抗性扰动。与 L-BFGS 相比，FGSM 只需要在反向传播阶段进行计算，因此其生成对抗样例的速度更快，更适合需要生成大量对抗样例的场景。假设 x 和 y 分别是原始图像和对应的标签。$J_{\#}$ (x,y)为损失函数。根据攻击行为的不同，将 FGSM 进一步分为针对性攻击和非针对性攻击。Goodfellow 等[55]对 FGSM 的攻击可表示为式(5)，Kurakin 等[81]将其扩展为针对性攻击，表示为式(6)。

$$x\ ?= x+\ \varepsilon \times 号(\nabla_x\ J_{\#}\ (x,y)),\ 非目标， \tag{5}$$

$$x\ ?= x-\ \varepsilon \times 符号(\nabla_x\ J_{\#}\ (x,y?)),\ 目标在\ y? \tag{6}$$

以针对性攻击为例。FGSM 首先识别概率最小的类作为目标。随后，从一组扰动中减去原始图像，从而生成一个足以误导模型输出目标类的对抗性示例。最近，一些研究对 FGSM 进行了改进，如 R+FGSM[149]和 MI-FGSM[39]。值得注意的是，FGSM 不仅在图像分类领域取得了显著的成就，而且在语音识别领域也取得了显著的成就。之前的研究已经探索了如何提取声学特征来生成音频对抗样例，Gong 等人[53]提出了第一个基于梯度符号的端到端音频对抗样例生成方法。

BIM 和我。FGSM 通过以非迭代的方式在梯度方向上的每个像素点上添加扰动来生成对抗性示例。相比之下，Kurakin 等人[**80**]的**基本迭代法(BIM)**迭代 FGSM (I-FGSM)使用迭代方法来查找每个像素点的扰动。BIM 的形式表示如下:

$$x'_0 = x, \quad x'_{i+1} = Clip_{x,\epsilon}\{x'_i + \alpha \times \mathrm{sign}\,(\nabla_x J\,(x'_i, y_{true}))\}. \tag{7}$$

每次，基于先前生成的对抗性示例 x?，单个像素以 α 的步骤增长(或减少)。然后执行裁剪操作 $Clip_{X,\varepsilon}$ {x?}，以约束新示例的单个像素不会偏离原始图像 x 太多。理想情况下，可以找到每个像素的变化小于 ε 的对抗性示例。在最坏的情况下，可以达到与 FGSM 相同的效果。此外，Kurakin 等人[80]设计了类迭代方法(Class Iterative Methods Method, ILCM)，将*对* $_{LLtrue}$ *抗性*扰动中的类 yin 替换为概率最低的类 y。

JSMA。Papernot 等人[114]仔细研究了模型输入特征与输出结果之间的关系，并介绍了 JSMA 对抗样例生成算法。它由以下三个主要过程组成:

(1)计算正导数∇ *F* (X)=∂ *F*∂ *x* (X)=[∂ *F*∂ *x*(*X*)]*i*∈1,*M*,*j*∈1,*N*。

(2)*使用*(1)的结果计算对抗性显著性 map。显著性 map 的值越高，表明相应的输入特征显著影响分类结果。

(3)根据(2)中的显著性图选择需要扰动的重要像素点，对所选像素点的像素值进行增减迭代，直到模型将其作为目标类输出。

JSMA 是隐形的，只需要干扰一些重要的像素点，而不是整个图像。此外，JSMA 已用于 NLP 和恶意软件检测。Grosse 等人[59]利用雅可比矩阵生成 Android 恶意软件的对抗性示例，以逃避基于 dnn 的恶意软件检测器的检测。

C&W。Carlini 和 Wagner[21]提出了一种基于优化的算法 C&W，通过设置一个度量输入和输出之差的损失函数来攻击防御性蒸馏网络。损失函数包含一个可调的超参数和一个控制对抗性样本置信度的参数，其可以正式表示为 $f(x?) = \max(\max\{Z(x?)_i : i ?t\} - Z(x?)_t, -\kappa)$。Z 表示 softmax 函数，$\kappa$ 为常数。C&W 算法根据范数分为?$_{20}$、?、?∞。

- ?$_0$ 意味着逐渐找到对分类结果影响较小的像素点，然后固定这些像素点。直到再也找不到这样不受影响的像素点，对剩下的像素点进行扰动。

- ?$_2$ 允许在修改的程度和数量之间取得平衡:

$$\text{最小化? ?12} \quad -\times(\tanh(w)+1)-x\Big\|_2^2 + c\times f\left(\frac{1}{2}\times(\tanh(w)+1)\right).$$

-对更改和迭代更新的?∞限制代替?$_2$ 惩罚:minimizec ⁄ f (x+ δ) + + $_i[(\delta_i-\tau)^+]$.

在 NLP 领域，Sun 和 Tang[143]提出了 C&W 算法，攻击病例预测模型，获取每个医患记录中的敏感内容，包括患者的病史信息和医生提出的治疗方案。

UAP。与 FGSM、DeepFool、ILCM 仅摄动单幅图像不同，**通用对抗摄动(Universal Adversarial 摄动，UAP)**是一种对多幅图像进行摄动的特殊方法。在多幅图像上添加通用摄动会导致图像被错误分类，而不解决优化问题或梯度计算[104]。具体来说，对于 *d* 维数据分布 μ，对于每个样本 $x\in Rd$，存在一个分类器 *k(x)*: k(*x* + δ)?K(x) μ 表示“大多数”x ~ μ。微扰 ξ *表示*?p 球的半径，满足约束 δ p≤ ξ。x ~ μ P (k(x + δ)?k(x))ˆ ≥1− *p*。的

目标是识别一个扰动向量 δ，它可以被添加到所有的例子中，并且这些例子被错误分类的概率为 1− *p*。请注意，不同的网络得到不同的扰动结果，甚至同一网络的不同初始化也不会得到相同的扰动结果。对 UAP 的进一步研究，如 GD-UAP[107]、NAG[125]、AAA[124]等。

DeepFool。FGSM 需要手动选择扰动因子 ε。为了解决这个问题，Fawzi 和 Frossard 提出了一种更具适应性的 DeepFool 攻击[105]和分类器鲁棒性的评估指标。DeepFool 不需要选择扰动因子 ε，并通过多个线性近似实现对一般非线性决策函数的攻击。对于二元分类器和线性分类器，可以使用超平面来区分这两类。我们只需要更新 ***x***

到直线或超平面的另一边得到 $x?_f(x)$的投影距离 $\frac{w}{\|w\|_2}$

最小摄动 θ 可以通过测量 x 到直线的最短距离，使 x 加上这个距离，然后转换到平面的另一边来求。Deepfool 既可以进行非目标攻击，也可以进行目标攻击。非目标攻击遍历所有类，找到变异最小的样本，而目标攻击则针对目标类的超平面执行[105]。

5.2.2 黑匣子攻击。

ATNs。大多数传统的对抗性示例都是基于梯度信息构建的，因此只适用于白盒攻击场景。相比之下，Baluja 和 Fischer 提出了两种对抗性示例生成方法**摄动 - atn(P-ATN)**和 **Adev- ersarial 自动编码 (AAE)**[**8**]，用于攻击基于**对抗性转换网络 (ATNs)**的一个或多个网络。ATNs 的优化目标是最小化联合损失函数 L_x 和 $LY_{\mathcal{U}}$生成对抗性示例:$\arg\min_\theta \sum_{xi\in X} \beta L_X(\Pi f, \theta(xi), xi) + LY(f(\Pi f, \theta(xi)), f(xi))$。ATNs 不仅可以进行目标攻击，也可以进行非目标攻击，而且还可以选择以黑盒或白盒方式训练网络。

ZOO。受 Carlini 和 Wagner[21]工作的启发，陈等人[26]提出了 ZOO 攻击，该攻击不需要模型的任何信息，通过估计目标模型的梯度直接生成对抗性示例。陈通过修改其损失函数提出了 ZOO 攻击。ZOO 攻击既可以进行目标攻击，也可以进行非目标攻击。在有针对性攻击的情况下，损失函数为式(5)，对于无针对性攻击，损失函数可替换为式(6)。

$$f(\mathbf{x},t) = \max\left\{\max_{i\neq t}\log[F(\mathbf{x})]_i - \log[F(\mathbf{x})]_t, -\kappa\right\}, \tag{8}$$

吗?吗?

$$f(\mathbf{x}) = \max\left\{\log[F(\mathbf{x})]_{t_0} - \max_{i\neq t_0}\log[F(\mathbf{x})]_i, -\kappa\right\}. \tag{9}$$

ZOO攻击和 C&W 攻击的性能相当，但 ZOO攻击查询和估计梯度的成本更高。

胡迪尼。胡迪尼是一种基于网络的可微损失函数的梯度信息产生对抗性扰动的方法[29]，可以针对任务损失进行定制。它不仅应用于图像分类，还应用于语音识别和语义分割。胡迪尼既可以用作非目标攻击，也可以用作目标攻击。在他们的论文中，作者描述了胡迪尼算法如何应用于三个不同的领域:人体姿势估计、语义分割和语音识别。ABX 测试结果表明，使用 DeepSpeech-2 模型[3]无法区分基于 houdini 的算法从原始音频中生成对抗性示例。

一个像素。Su 等人[141]执行了一种称为一像素攻击的对抗性攻击算法，该算法只需要改变少量像素点。在极端情况下，只需要修改一个像素点，不需要修改网络参数或梯度的信息来进行攻击。单像素攻击也适用于梯度不可微或难以计算的攻击模型。单像素攻击对扰动的大小没有太多限制。扰动被编码成矩阵，即候选解。应用微分进化来获得最优解。一个候选解包含固定数量的扰动。每个扰动包含 5 个分量，x、y 坐标和扰动 RGB。每次修改一个像素。首先，对每个像素进行迭代和修改，生成一个子图像

(候选解)，随后与母图对比。根据预设的判定规则，相应保留攻击效果最好的子图像，并进行下一次迭代。最后，当达到预设的迭代次数，或者实际类标签低于 5%时，停止迭代，对抗性攻击完成。

EvadeML。Xu、Qi 和 Evans[167]使用遗传编程攻击**可移植文档格式(Portable Document Format,PDF)**恶意软件分类器。变体的突变由检查运行时行为的杜鹃沙盒提供的分数控制。在他们的工作中，遗传编程中的每个变异都由目标分类器和神谕器进行评估。目标分类器是一个黑盒子，输出恶意软件的置信度评分。接下来，被选中的变体被突变操作员随机操纵，以产生下一代种群。这个过程会一直持续下去，直到找到一个可规避的样本，或者达到一个阈值的代数。作者报告了针对 PDFrate 和 Hidost 恶意软件分类器的 100%逃避率，但承认该过程在计算上是昂贵的。

MalGAN。最近，引入了 MalGAN[69]来生成 PE 可执行恶意软件对抗示例。该方法创建了一个完全可微的替代模型，经过训练后可以生成带有相应输入的修改后的恶意软件特征。然后将替代模型用于改进的 GAN 中的梯度计算，以产生规避的恶意软件变体。生成器用于生成对抗性示例。恶意软件特征向量 ***m*** 和噪声向量 ***z*** 是生成器网络的输入。***m*** 中的每个元素表示一个特征是否存在。噪声向量 ***z*** 从均匀分布[0,1)中随机采样，在恶意软件特征生成过程中从二进制恶意软件中删除特征可能会导致故障。因此，作者选择添加不相关的特征，而不是删除它们。MalGAN 算法不依赖于模型梯度，具有较高的攻击成功率。但是，与梯度攻击方法相比，它需要更大的开销。

5.2.3 其他攻击。

一语音识别。CommanderSong[171]设计了第一个针对 ASR 系统的鲁棒跨空实用对抗性攻击。研究人员在物理攻击中考虑了扬声器噪声、接收机失真和通用背景噪声，并首先建立了噪声模型。随后，在歌曲 *X(t)*中加入噪声和随机小扰动，然后使用梯度下降优化使修改后的歌曲 *X?(t)*的 pdf-id 与命令 *b* 相似，从而愚弄 Kaldi[*117*]。Alzantot 等人[2]阐述了第一个基于遗传算法的黑盒攻击音频对抗实例生成，从而避免了计算 MFCC 衍生物的繁琐工作。攻击者通过注入随机噪声创建一个候选对抗实例的种群，然后计算每个种群成员的适应度得分，通过应用选择、交叉和突变，从当前一代中生成下一代对抗实例。受 Alzantot 的工作[2]的启发，Taori 等人[148]基于遗传算法和梯度估计制作了有针对性的对抗性音频。为了使其适合短语和句子，引入了动量突变和 CTC Loss。实验表明，他们的研究可以在更复杂的深度语音系统上解码任何长度的短语。

-自然语言处理。基于维基百科的阅读理解示例，Jia 和 Liang 执行了 ADDSENT 和 ADDANY，生成了看起来与问题相似但在语义上与正确答案不一致的句子[73]。Song 和 Shmatikov[136]发现，对抗性样本同样能够欺骗**光学字符识别(OCR)**。因此，他们生成了个别单词的印刷文本的对抗性图像，足以导致 Tesseract 系统错误识别主体

表 6 所示。对抗性例子防御方法

防御动机	防御策略	防御的方法	优势	缺点
模型	修改网络	模型防御蒸馏[115]	易于训练，开销低更好泛化，能力强	与模型高度相关，高计算复杂度
		梯度正则化[106]	模型鲁棒性增强	
		深度压缩网络[60]	高鲁棒性	
	模型改进	特征压缩[166]	低复杂度	降低模型预测性能
	额外的网络	探测器[95]	低复杂性，模型无关的	更糟糕的泛化，对模型稳健性没有贡献
		基于 gan 的[74,82]		
	验证网络	网络验证[56,75]	高可解释性	巨大的计算开销
数据	修改培训过程或输入数据	对抗性训练[1,55,129,144,149]	容易实现，防御能力强	难以收敛，高计算开销
		数据随机化[163]		
		数据压缩[41]		
		输入重构[36,138]		

用反义词代替原词的。文本数据的离散性强化了对 NLP 任务的对抗性示例攻击。目前仍有大量的作品克服了这样的困难。为了解决这一挑战，Li 等人[83]提出了 Bert-Attack，通过替换具有高语义影响的单词来确保生成的对抗性样本的语义一致性和句子流畅性。

-恶意软件检测。Gross 等人[59]对 Android 恶意软件检测中的恶意软件二进制特征进行了对抗性示例攻击。Hu 等人[68]演示了一种对抗性示例序列生成方法，该方法可用于攻击 RNN 检测系统中的模型。然而，Arjovsky 和 Bottou[5]指出，GAN 存在与训练相关的稳定性问题，并且它可能不会在给定的数据集中收敛，[68]中的攻击方法也会遇到这样的问题。Anderson 等人[4]首次采用强化学习来逃避基于 ml 的恶意软件检测。具体来说，他们预先定义了一系列功能无关的修改操作，然后进行强化学习以获得最优的修改操作序列，但攻击的成功率仅为 24%。

5.3 防御

对抗性防御技术应避免对原始模型结构的过度修改，同时确保模型的性能(如速度、内存使用、模型分类精度)不被削弱。针对对抗性示例攻击的防御方法在图像分类领域得到了广泛的研究，语音识别和恶意软件领域次之，NLP 领域最少[20]。因此，我们将重点阐述针对图像分类和其他几个领域的对抗性示例攻击的防御方法。我们在表 6 中总结了防御方法。

*5.3.1 防御性蒸馏。*Hinton 等人[65]首次报道了对蒸馏的系统研究。蒸馏背后的思想是将细粒度的知识从大规模训练模型迁移到小规模模型，使小规模模型能够更准确、更高效地执行学习任务。Papernot 等人[115]在 Hinton 研究的基础上，批准了一种防御性蒸馏方法。提供一个小模型(蒸馏网络)来模拟一个大的、计算密集型的模型(初始)，在不影响准确性的情况下解决信息缺失的问题。防御性蒸馏能够防御大多数对抗性示例攻击，并且易于训练。如图 13 所示，具体防御过程如下:首先，使用硬标签训练教师模型/初始网络，设置温度 t，初始网络的 softmax 层的输出类概率为 $F(X) = [e_{y]-1} e^{zl(X)/T} \in_i 0\cdots N-1$。接下来是交叉熵 $zi^{(X)/T}$

$l=0$

《ACM 计算调查》，第 55 卷第 1 期，第 20 条。出版日期:2021 年 11 月。

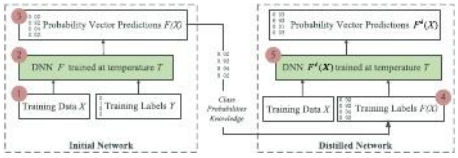

图 13 所示。防御性蒸馏的架构[115]。

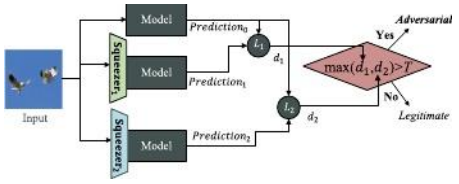

图 14 所示。检测对抗性样本的特征压缩框架[166]。

计算损失函数。然后利用从教师模型中得到的软标签来训练蒸馏的网络(学生模型)。最后，对蒸馏网络的最后一层(学生模型)进行修改，并设置温度 T= 1，从而以高置信度预测未知输入的类别。防御性蒸馏的优点是泛化能力高，训练开销低。然而，它对神经网络的鲁棒性没有贡献[19,21]。

5.3.2 梯度正则化(Gradient Regularization)。梯度正则化是指在训练过程中对目标函数添加约束，以避免模型输出随着输入的变化而发生显著变化。通常情况下，小的扰动不会显著影响输出。Lyu 等人[21]使用一组联合的正则化方法训练模型来防御基于 fgsm 的攻击。值得注意的是，对抗性训练显著降低了损失函数的曲率和分类器的决策边界。相应地，Moosavi-Dezfooli 等人[106]提出了一种新的曲率正则化策略，直接最小化损失曲面的曲率。他们的方法显著提高了神经网络的鲁棒性，但可能会损害模型的性能(例如，降低精度)。此外，正则化方法和对抗性训练可以结合起来防御对抗性攻击，但计算复杂度太高。

5.3.3 深度压缩网络。Gu 和 Rigazio 基于**收缩自动编码器(CAE)**构造了深度收缩网络**(DCN)**，并证明了该方法有效地提高了神经网络的鲁棒性[60]。他们训练**去噪自动编码器(DAE)**来去除对抗性噪声。DCN 的目标函数由自编码器和平滑惩罚组成，是一个端到端的训练过程。$J_{DCN}(\theta) = \sum_{i=1}^{m} L(t^{(i)}, y^{(i)}) + \lambda \sum_{j} \left\| \frac{\partial h_j^{(i)}}{\partial h_{j-1}^{(i)}} \right\|_2$，其中 $t^{(i)}$和 $y^{(i)}$表示输入 $x^{(i)}$的真实和预测标签，re-

spectively。惩罚项由比例因子 λ 和 *Frobenius* 范数组成

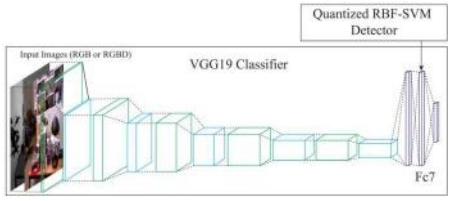

图 15 所示。SafetyNet 由传统分类器和 RBFSVM 组成[95]。

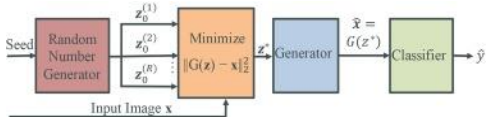

图 16 所示。防御 GAN 算法概述[128]。

不修改模型本身，因此可以很容易地与其他防御策略集成，以获得更好的防御结果。

5.3.5 探测器。检测器是一种区分检测到的图像是否为对抗性样本的机制。通常，检测器区分对抗性样本的标准可以自由定义。最直接的方法是标记对抗性和合法样本来训练分类器。分类器训练方法主要有两种。一种方法是通过在初始阶段直接将对抗样本和原始样本分开标记来训练分类器。另一种方法是通过仅在特定层的输出值上标记对抗性样本和干净样本来训练分类器。Metzen 等人[103]基于一个小型检测器完成了一个二元分类任务的训练，该检测器用于区分真实数据和对抗样本。Li 等人[84]总结了恶意软件检测领域的各种类型的对抗性攻击，并使用被操纵的数据集训练了一个 SVM 分类器来防御对抗性示例攻击。如图 15 所示，Lu 等人[95]开发了 SafetyNet，其中 Relu 函数的输出用作检测器的特征，并通过 RBF-SVM 分类器区分合法样本和对抗样本。SafetyNet 由原始分类器(VGG19/ResNet)和对手检测器(RBF-SVM)组成，如果检测器显示它们是敌对的，则拒绝样本。此外，研究人员还基于 PCA[87]、最大平均差异[58]、不确定性[46]等标准来区分对抗性样本。

5.3.6 Gan-Based。由于 GAN 的兴起及其优异的性能[128]，许多研究人员将其应用于对抗性示例防御[74,82]。Samangouei 等人[128]降低了基于 GAN 的对抗性扰动的效率，并提出了一种防御策略- GAN，一种既能防御白盒攻击又能防御黑盒攻击的策略。如图 16 所示，Defense-GAN 不修改分类器结构和训练过程。然而，训练一个稳定的 GAN 网络和选择合适的超参数是至关重要的，否则防御 GAN 无法达到预期的防御效果。具体过程如下:

(1)生成 R 个随机噪声向量 $\mathbf{z0}_{(1)}$， $\mathbf{z0}_{(2)}$，…，基于随机数生成器生成 $\mathbf{z0}_{(R)}$。

(2)馈入随机向量 $\mathbf{z}_{0(1)}$， $\mathbf{z}_{0(2)}$，…， $\mathbf{z}_{0(R)}$和样本 $\mathbf{x}$ 到训练好的 GAN 网络中，以找到满足目标函数最小化 $G(\mathbf{z}) - \mathbf{x}_2^2$ 的 z^*。

(3)将 $\mathbf{z}^*$ 送入生成器网络进行训练，直到生成器生成满足干净样本分布的图像 $\mathbf{x}^*$。

(4)分类 $\mathbf{x}' = G(\mathbf{z}^*)$。

5.3.7 网络验证。网络验证可用于检查样本是否违反了 dnn 的某些属性，或者在确定范围内(与原始样本的距离)的样本是否改变了其标签值。在对抗性示例攻击的防御阶段，如果检测到一个对抗性示例的输入违反了 dnn 的某些属性，则对其进行区分。

因此，网络验证是基于模型本身对新的未知攻击的检测。Katz 等[75]设计了基于可满足模理论解算器的 Reluplex 检测方法来验证神经网络的 Relu 激活函数。神经网络验证的计算复杂度较高，属于 np 完全问题，因此 Katz 通过对待检查节点的顺序进行优先排序和共享验证信息来加快验证速度。Reluplex 算法只检查几个单独输入点附近的鲁棒性。Gopinath 等人[56]对 reluplex 方法进行了扩展，提出了 Deepsafe。它自动识别和验证输入空间的安全区域，其中网络对特定标签错误分类具有鲁棒性，并且可以防御特定的目标攻击。

*5.3.8 对抗性训练。*顾名思义，对抗性训练意味着在训练阶段将攻击算法(如 FGSM)制作的对抗性示例添加到训练集中。它是一种蛮力防御方案，也是一种缓解模型过拟合问题的正则化工具。对抗性训练需要大量的对抗性示例来训练网络对抗单步攻击，但对迭代攻击无效。对抗性训练的第一批研究来自 Goodfellow 等人[55]，他们通过这种方式获得了一个更鲁棒的模型。这也意味着模型防御对抗性样本的能力得到了提高。他们发现，这种方法可以很好地防御白盒攻击，但在黑盒场景中并不有用。Tramer 等人[149]进一步进行了集成对抗训练，扩大了训练集，从而增加了对抗样本的多样性，即使在黑箱场景下也取得了更好的防御效果。在音频领域，Sun 等人[104]首先提出在 FGSM 动态生成的自然样本补充对抗样本上训练鲁棒声学模型。他们在两个不同的数据集上验证了该方法:Aurora-4、CHiME-4，发现对抗性训练有效地提高了卷积神经网络的鲁棒性。然而，无论防御方式如何，总会发现新的对抗性示例发动对抗性攻击[104]。

*5.3.9 数据随机化。*数据随机化这个术语是指执行随机化技术来减轻对抗性效应。Xie 等人[163]通过随机调整图像大小并使用随机填充技术在模型的前向传播阶段破坏特定对抗性扰动的结构来减轻对抗性效应。具体来说，它由两个步骤组成。(i)首先，随机调整输入图像的大小。(ii)然后，在调整大小的图像周围应用零随机填充。填充的位置是随机选择的。数据随机化在防御单步攻击和迭代攻击方面都是有效的。数据随机防御方法不仅不需要额外的训练，计算强度也较低，而且还与其他对抗性防御方法兼容。

*5.3.10 输入重建。*Song 等人[138]发现摄动和良性图像之间的对数似然分布存在显著差异。在图像训练集分布中，对抗性样本的概率密度远低于良性样本，并且主要位于低概率区域。在此基础上，Song 等人[138]提出了 PixelDefend 防御方法，通过重构的方式对输入数据进行净化，将其移回训练分布的高概率区域，净化恶意对抗的图像。具体来说，对于输入图像 **X**，目标是找到在用于控制示例重建行为的 **X.**ϵ is_{defend} 的 ϵ -ball 内 $_{\text{defend}}$ 分布 P(X)的概率最大化的图像 X *。如果没有检测到对抗性样本，则不改变样本，$\lambda = O_{\text{defend}}$ 。PixelDefend 防御方法不需要对模型进行重新训练，但可能不适用于数据集空间分布较大的场景。

*5.3.11 其他防御。*生成对抗性示例的方法多种多样，很难使用一种防御技术来抵抗所有对抗性示例攻击。因此，许多研究人员开始研究各种防御技术的组合，以增强防御的有效性。例如，Meng 等人[101]建立了 MagNet 防御框架，该框架将一个或多个独立的检测器网络和一个重整器网络结合在一起，既能防御黑盒攻击，也能防御灰盒攻击。Raghunathan 等人[122]引入了一种基于半确定松弛的技术来生成证书。然后使用这些证书来训练一个更健壮的网络，以对抗对抗性示例攻击。Prakas 等人[118]通过结合像素偏转和小波去噪技术，执行了一种新的集成防御方法，该方法在小波域中使用自适应软阈值来平滑模型的输出。这种防御方法可以有效防御最新的对抗性攻击。Zhang 等人[176]使用**电压过缩放(Voltage Over Scaling,VOS)**技术在随机数据上训练了一个多冗余架构防御模型，以防御对抗性示例攻击。由于他们的防御方法 HRAE 需要大量的训练，所以它更适合于离线环境。梯度掩蔽试图隐藏网络的梯度信息，从而防止攻击者使用梯度求解方法构造对抗性示例。然而，梯度掩蔽很难防御黑盒攻击[149]。

6AI 系统集成

AI 技术无处不在。虽然我们已经讨论了 AI 本身的威胁和对策，但当 AI 集成到现实世界的应用程序中时，安全问题似乎更加复杂。对于不同的应用场景，安全问题是不同的，我们应该从全局的角度来看待 AI 的安全。本节将探讨实际集成阶段的几个安全风险。

-AI 机密性。AI 的机密性包括数据机密性和模型机密性。AI 的机密性通常与模型隐私明确相关，尽管它也可能(间接)导致安全问题[175]，如模型反演[48]和模型提取[150]。模型反演是指基于输入和输出之间的映射关系，对模型进行逆分析，以获得私有数据。另一方面，模型提取通常被理解为通过 API 执行可接受数量的查询，并观察输出结果(概率或标签)来推断模型参数或提取与目标模型密切匹配的近似模型。对于这两种类型的隐私问题，通常使用 dp 差分隐私[40,44]、同态加密[52]或模型水印[153]来减轻隐私风险。目前的 AI 安全和隐私问题似乎是分开解决的。尽管具有挑战性，但值得考虑系统地、并发地解决这些问题[72,137]，以确保数据和模型隐私，同时保持 AI 系统的安全性。

由 Google 研究团队提出的联邦学习[98]是一种新的分布式机器学习技术，已成为 AI 的另一个重要新分支。联邦学习聚合了在每个客户端持有的本地化数据上训练的本地模型，以更新全局模型。联邦学习主要缓解了隐私问题，但它缺乏对本地数据的审计和对参与者行为的控制，这很可能会引入安全问题。联邦学习的客户端、中央服务器和通信渠道很容易成为攻击者的目标。最典型的是投毒攻击，可以通过数据投毒或/和模型投毒来实现[7,11]，用户端/服务器端 GAN 攻击[157]，以及联邦学习中的隐私问题[63]。由于缺乏对数据的访问权限和对客户端的控制有限，设计针对安全攻击的对策比集中训练更具挑战性。

—代码漏洞保护。当前的 AI 系统技术(如深度学习)是建立在框架上的(如 Tensorflow、Caffe 和 Torch)。这些框架依赖于各种基础库和第三方组件，这些组件极大地促进了 AI 技术的发展。然而，它们并不是设计得完美无缺，存在漏洞[162]。更核心的是，最近的一项研究表明，未经审计的第三方代码片段——损失计算的代码响应——可以在深度学习模型中植入后门[6]。虽然代码漏洞是模型实现过程的一部分，但它们也是 AI 系统安全部署的关键部分。

7 个挑战与机遇

虽然上面提到的攻击和对策已经被广泛研究，但仍然存在各种挑战和机遇。

-确定一种可以应用于不同类型传感器设备的防御机制。由于新的数据采集设备和工具的不断更新，在数据采集过程中，保护数据的安全性还有很大的进步空间。传感器的安全问题主要在于硬件的设计以及攻击者对硬件进行的恶意信号注入。在硬件方面，研究人员需要对传感器的物理特性和逻辑布线进行修改和验证。在软件方面，研究人员需要设计合理的策略，在传感器的数据采集阶段识别和拒绝恶意信号，比如传感器需要对数据信号进行清洗、放大等预处理操作。其次，传感器恶意信号识别技术通用性差。迫切需要识别一种可以应用于不同种类传感器设备的防御机制，使攻击者无法绕过防御机制。

-增强 AI 模型的可解释性。目前对 AI 系统的研究缺乏模型的可解释性。然而，我们很难明确地分析和解释 AI 在什么情况下会出现什么安全问题。因此，攻击者总是能够找到 AI 的攻击面，而我们无法在攻击者暴露和利用之前识别潜在的安全漏洞。我们迫切需要加强 AI 模型的数学验证，增强 AI 的可解释性。

——加强 AI 系统的安全和隐私双重保护。安全和隐私是密切相关的，AI 系统不仅面临上述安全挑战，还面临隐私风险[175]，如模型反演[48]和模型提取[150]。目前 AI 的安全和隐私问题是分开解决的。未来需要加强隐私和安全领域的融合研究[72,137]，以确保数据和模型的隐私问题，同时保持 AI 系统的安全性。

-整体安全保障(Holistic Security Assurance)。正如 Bertino[10]所指出的，我们可以使用整体安全保障来减轻 AI 的安全威胁。简单地专注于特定事件，例如对抗性示例攻击或数据中毒，是不够的或无效的。这可能会忽略一些事实。例如，虽然防止对交通标志图像的对抗性示例攻击很重要，但利用多个信息源(如地图标记和雷达信息)共同做出决策，消除单一来源的风险是值得的。此外，在集成 AI 应用程序时，可以综合考虑系统安全性、数据安全性和软件安全性，构建一个整体的保障流程。也就是说，在其他安全领域的实践，例如数据安全中的数据来源，可以有效地降低数据中毒攻击的风险，有助于构建安全的 AI 系统。

8 的结论

AI 已广泛应用于各个领域，将引发新一轮产业变革，推动人类社会进入智能时代。然而，AI 技术的发展尚未成熟。安全风险存在于 AI 系统生命周期的各个阶段。尽管已经提出了许多对策，但仍有各种挑战需要解决。本文详细介绍了 AI 系统在数据收集、模型训练、模型推理和集成应用等方面的安全研究进展，并对相关防御技术进行了综述。最后，我们总结了与 AI 系统相关的安全问题的挑战和机遇。虽然 AI 发展技术如火如荼，但其日益突出的安全问题也促使我们更加警惕，保护其更高、更快、更好的发展。

参考文献

[1] Abdullah Al-Dujaili, Alex Huang, Erik Hemberg，和 Una-May O 'Reilly。2018.用于稳健检测二进制编码恶意软件的对抗性深度学习。*2018 年 IEEE 安全与隐私研讨会论文集*。IEEE 76 - 82。

[2] Moustafa Alzantot, Bharathan Balaji, Mani B. Srivastava. 2018。你听到了吗?对抗自动语音识别的对抗性例子。*CoRR abs/1801.00554(2018)。arXiv:1801.00554*。http://arxiv.org/abs/1801.00554。

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai，白敬良，Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro，程强，陈国梁，陈杰，陈京东，陈志杰，Mike Chrzanowski, Adam Coates, Greg Diamos，丁柯，杜牟东，Erich Elsen, Jesse Engel，方伟伟，樊林西，Christopher Fougner，高亮，龚采霞，Awni Hannun, Tony Han, Lappi Vaino Johannes，蒋兵，蔡菊，Billy Jun, Patrick LeGresley, Libby Lin，刘俊杰，刘洋，李伟高、李祥刚、马东鹏、Sharan Narang、Andrew Ng、Sherjil Ozair、彭一平、Ryan Prenger、钱盛、权宗锋、Jonathan Raiman、Vinay Rao、Sanjeev Satheesh、David Seetapun、Shubho Sengupta、Kavya Srinet、Anuroop Sriram、唐海远、唐丽良、王冲、王继东、王开复、王怡、王志坚、王志谦、吴爽、魏丽凯、肖博、谢文、谢焱、丹尼 Yogatama、袁斌、詹俊、朱振耀。2016。深度语音 2:英语和普通话的端到端语音识别。*《第 33 届国际机器学习国际会议论文集》*，173-182。

[4] Hyrum S. Anderson, Anant Kharkar, Bobby Filar, and Phil Roth。2017。规避机器学习恶意软件检测。*黑帽*(2017)，1-6。

[5] Martin Arjovsky and l<s:1> Bottou. 2017。《走向训练生成对抗网络的原则方法》。arXiv:1701.04862 [stat.ML] https://arxiv.org/abs/1701.04862

[6] Eugene Bagdasaryan 和 Vitaly Shmatikov. 2020。深度学习模型中的盲后门。CoRR abs/2005.03823(2020)。arXiv:2005.03823 https://arxiv.org/abs/2005.03823

[7] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, Vitaly Shmatikov. 2020。如何借壳联邦学习。*《第 23 届国际人工智能与统计会议论文集》*。PMLR，2938 - 2948。

[8] Shumeet Baluja 和 Ian Fischer. 2017。对抗性转换网络:学习生成对抗性示例。CoRR abs/1703.09387(2017)。arXiv:1703.09387 http://arxiv.org/abs/1703.09387

[9] Mauro Barni、Kassem Kallas、Benedetta Tondi. 2019。cnn 中一种新的无标签中毒训练集腐败的后门攻击。*《2019 年 IEEE 图像处理国际会议论文集》*。IEEE 101 - 105。

[10] E. Bertino. 2021。对人工智能的攻击[最后一句话]。*IEEE 安全与隐私* 19,01 (Jan 2021)， 103-104。DOI:https://doi.org/10.1109/MSEC.2020.3037619

[11] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. 2019。通过对抗性镜头分析联邦学习。*《机器学习国际会议论文集》*。PMLR，634 - 643。

[12] Bryan Biegel, James F. Kurose. 2016。国家人工智能研究与发展战略规划_nstc 和 NITRD。白宫(2016)。https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf。

[13]巴蒂斯塔·比乔、伊吉诺·科罗娜、乔治·富梅拉、乔治·贾辛托、法比奥·罗利。2011。对抗性分类任务中对抗投毒攻击的 Bagging 分类器。在*多重分类器系统*中。C. Sansone, J. Kittler 和 F. Roli(主编)，《计算机科学课堂笔记》(包括人工智能和生物信息学的子系列课堂笔记)，第 6713 卷，Springer, 350-359 页。DOI:https://doi.org/10.1007/978-3-642-21557-5_37

[14]刘建军，刘建军。2013。针对支持向量机的中毒攻击。arXiv:1206.6389 [c .lg]。

[15]巴蒂斯塔·比乔和法比奥·罗利。2018。野性模式:对抗性机器学习兴起十年后。*2018 年 ACM SIGSAC 计算机与通信安全会议论文集(加拿大多伦多)(CCS' 18)*。美国计算机械协会，纽约，NY, USA, 2154-2156。https://doi.org/10.1145/3243734.3264418

[16] Benjamin Birnbaum, Brian DeRenzi, Abraham D. Flaxman, Neal Lesh。2012。移动数据采集的自动化质量控制。*第二届ACM计算促进发展研讨会论文集*。ACM，纽约，NY。DOI:https://doi.org/10.1145/2160601.2160603

[17] Jakramate Bootkrajang and Ata Kabán。2014. 在类标签噪声存在的情况下学习核逻辑回归。*模式识别* 47,11(2014)，3641-3655。DOI:https://doi.org/10.1016/j.patcog.2014.05.007

[18] Tom B. Brown, Dandelion man<e:1>， Aurko Roy, Martín Abadi, and Justin Gilmer. 2017。敌对的补丁。CoRR abs/1712.09665(2017)。arXiv:1712.09665 http://arxiv.org/abs/1712.09665

[19] Nicholas Carlini 和 David A. Wagner. 2016。防御性蒸馏对对抗性例子不鲁棒。CoRR abs/1607.04311(2016)。arXiv:1607.04311 http://arxiv.org/abs/1607.04311

[20] Nicholas Carlini and David Wagner. 2017。对抗性示例不容易被检测:绕过十种检测方法。*第10届ACM人工智能与安全研讨会论文集*，3-14。

[21] Nicholas Carlini 和 David Wagner. 2017。走向评估神经网络的鲁棒性。*2017年IEEE安全与隐私研讨会论文集*。IEEE，39-57。

[22] Nicholas Carlini, David Wagner. 2018。音频对抗性示例:针对语音到文本的针对性攻击。*2018年IEEE安全和隐私研讨会论文集*，1-7。

[23] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay 和 Debdeep Mukhopadhyay。2018。对抗性攻击和防御:调查。CoRR abs/1810.00069(2018)。arXiv:1810.00069 http://arxiv.org/abs/1810。00069

[24]纪尧姆 M. J.-B.。Chaslot, Mark H. M. Winands, H. Jaap van den Herik, Jos W. H. M. Uiterwijk和 Bruno Bouzy, 2008。蒙特卡洛树搜索的渐进式策略。*新数学与自然计算* *04,03*(2008年11月)，343-357。DOI:https://doi.org/10.1142/s1793005708001094

[25] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, Biplav Srivastava. 2018。基于激活聚类的深度神经网络后门攻击检测。CoRR abs/1811.03728(2018)。arXiv:1811.03728 http://arxiv.org/abs/1811.03728

[26]陈品宇，张欢，Yash Sharma，易金峰，谢卓瑞。2017。Zoo:不训练替代模型的基于零阶优化的深度神经网络黑盒攻击。*第十届ACM人工智能与安全研讨会论文集*，15-26。

[27]陈新云，刘畅，李波，宋黎明，2017。利用数据中毒对深度学习系统的针对性后门攻击。CoRR abs/1712.05526(2017)。arXiv:1712.05526 http://arxiv.org/abs/1712.05526

[28] Andreas Christmann and Ingo Steinwart. 2004。关于凸风险最小化方法在模式识别中的鲁棒性。The*Journal of Machine Learning Research*(2004年12月)，1007-1034。

[29] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet, 2017。胡迪尼:愚弄深度结构化预测模型。arXiv:1707.05373。

[30] Camille Cobb, Samuel Sudar, Nicholas Reiter, Richard Anderson, Franziska Roesner, and Tadayoshi Kohno. 2016。数据收集技术的计算机安全。*第八届信息和通信技术与发展国际会议论文集(Ann Arbor, MI，USA)(ICTD ' 16)*。计算机协会，纽约，纽约，USA，第2条，11页。https://doi.org/10.1145/2909609.2909660

[31] D. Cockburn 和 N. R. Jennings. 1996。ARCHON:用于工业应用的分布式人工智能系统。见*《分布式人工智能基础》*。G. M. P. O 'Hare和 N. R. Jennings(主编)，Wiley，319-344

[32]罗南·科洛伯特、克里斯蒂安·普赫施和加布里埃尔·辛纳夫。2016。Wav2Letter:基于端到端convnet的语音识别系统。CoRR abs/1609.03193(2016)。arXiv:1609.03193 http://arxiv.org/abs/1609.03193。

[33] Jamie Condliffe. 2015。自2000年以来，机器人手术与144例美国死亡有关。2021年10月28日检索自https://gizmodo.com/robotic-surgery-has-been-connected-to-144-u-s-deaths-s-1719202166。

[34]刘建军，刘建军，刘建军，等。2008。驱魔:异常传感器训练数据的消毒。*2008年IEEE安全与隐私研讨会论文集*。IEEE 81 - 95。

[35] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma. 2004。敌对的分类。*第十届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，99-108。

[36] Nilaksh Das, Madhuri Shanbhogue，陈尚 tse, Li Chen, Michael E. Kounavis, Duen hong Chau。2018。慢板:音频对抗性攻防的互动实验。《数据库中的*机器学习和知识发现欧洲联合会议论文集*》。施普林格,677 - 681。

[37]杰弗里·达斯汀。2018。亚马逊放弃了对女性有偏见的秘密AI招聘工具。检索于2021年10月28日，来源:https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai- recruit-tool-that-show-bias -against-women- iduskcn1mk08g。

[38]张建军，张建军，张建军，张建军，张建军，张建军，张建军。AI系统中大数据的隐私和安全:研究和标准视角。*《2019年IEEE大数据国际会议论文集》*。IEEE 5737 - 5743。

ACM Computing Surveys, Vol. 55, No. 1, Article 20。出版日期:2021 年 11 月。

[39]董银鹏，廖方舟，庞天宇，苏航，朱军，胡，李建国。2018。以势头助推对抗性攻击。*《IEEE 计算机学会计算机视觉与模式识别会议论文集》*。IEEE 9185 - 9193。DOI:https://doi.org/10.1109/CVPR.2018.00957

[40]辛西娅·德沃克。2006。微分隐私。在 *Automata, Languages and Programming* 中。M. Bugliesi, B. Preneel, V. Sassone 和 I. Wegener(主编)，《计算机科学课堂笔记》(包括《人工智能课堂笔记》和《生物信息学课堂笔记》子系列)。第4052卷，Springer, 1-12页。DOI:https://doi.org/10.1007/11787006_1

[41] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. 2016。JPG压缩对对抗性图像的影响研究。CoRR abs/1608.00853(2016)。arXiv:1608.00853 http://arxiv.org/abs/1608.00853

[42]比尔·艾德森。2020。Mitre、微软和其他11个组织对机器学习威胁进行了研究。检索于2021年10月28日，摘自 https://www.mitre.org/publications/project-stories/mitre-microsoft- others-take- machine-learning-threats。

[43] Paul Triolo Graham Webster, Rogier Creemers, and Elsa Kania, 2017。《下一代人工智能发展规划:中国》。检索于2021年10月28日，摘自 https://www.newamerica.org/cybersecurity-initiative/ digichina/blog/full- transla-chinas -new-generation-artificial-intelligence- develop-plan -2017/。

[44] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014。RAPPOR:随机聚合可保护隐私的有序响应。*《ACM 计算机与通信安全会议论文集》*。DOI:https://doi.org/10.1145/2660267.2660348

[45] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017。机器学习模型的鲁棒物理世界攻击。CoRR abs/1707.08945(2017)。arXiv:1707.08945 http://arxiv.org/abs/1707.08945

[46]鲁本·费曼，Ryan R. Curtin, Saurabh Shintre, Andrew B. Gardner. 2017。从人工制品中检测对抗性样本。arXiv:1703.00410 [stat.ML]。

[47]金融稳定委员会，2017。金融服务中的人工智能和机器学习——市场发展和金融稳定影响。*金融稳定委员会* 45(2017)。

[48] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, Thomas Ristenpart. 2014。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。第23*届USENIX 安全研讨会论文集*，17-32。

[49]高骥，Jack Lanchantin, Mary Lou Soffa，齐彦军。2018。对抗文本序列的黑盒生成，以逃避深度学习分类器。*2018 年 IEEE 安全与隐私研讨会论文集*。IEEE, 50-56。

[50]高岩松，段宝嘉，张智，马思琪，张纪良，傅安民，Surya Nepal, Hyoungshick Kim。2020。深度学习的后门攻击与对策:综合综述。CoRR abs/2007.10760(2020)。arXiv: 2007.10760 https://arxiv.org/abs/2007.10760

[51]高岩松，徐变，王德睿，陈世平，Damith C. Ranasinghe, Surya Nepal. 2019。Strip:一种针对深度神经网络的木马攻击防御。*第35届计算机安全应用年会论文集*，113-125。

[52]刘建军，刘建军，刘建军。2008。crypp - tonets:将神经网络应用于高吞吐量和准确性的加密数据。*机器学习国际会议论文集*，201-210。

[53]袁工和 Christian Poellabauer。2017。为语音副语言学应用制作对抗性示例。CoRR abs/1711.03280(2017)。arXiv:1711.03280 http://arxiv.org/abs/1711.03280

[54]龚志涛，王文录，李波，宋黎明，顾伟信。2018。基于梯度方法的对抗性文本。arXiv:1801.07175 [c]。CL] https://arxiv.org/abs/1801.07175

[55] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015。对抗性例子的解释和利用。*《第三届国际学习表征会议论文集》*。ICLR。

[56] Divya Gopinath, Guy Katz, Corina S. pere surireanu, Clark Barrett. 2018。Deepsafe:一种评估神经网络鲁棒性的数据驱动方法。*《验证与分析自动化技术国际研讨会论文集》*。施普林格,3-19。

[57] Alex Graves, Santiago Fernández, Faustino Gomez 和 j<s:1> rgen Schmidhuber. 2006。连接主义时间分类:用递归神经网络标记未分割的序列数据。*第23届机器学习国际会议论文集*，369-376。

[58] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick D. McDaniel。2017.关于对抗性样例的(统计)检测。CoRR abs/1702.06280(2017)。arXiv:1702.06280 http://arxiv.org/abs/ 1702.06280

[59] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, Patrick McDaniel。2017.恶意软件检测的对抗性示例。摘自*《欧洲计算机安全研究研讨会论文集》*。施普林格,62 - 79。

[60]顾世祥和卢卡·里加齐奥。2015。走向对对抗性示例具有鲁棒性的深度神经网络架构。arXiv:1412.5068 [c]。LG] https://arxiv.org/abs/1412.5068

[61]顾天宇，Brendan Dolan-Gavitt, and Siddharth Garg. 2017。坏网络:识别机器学习模型供应链中的漏洞。CoRR abs/1708.06733(2017)。arXiv:1708.06733 http://arxiv.org/abs/1708.06733

[62]顾天宇，刘康，Brendan Dolan-Gavitt, Siddharth Garg. 2019。坏网络:对深度神经网络的后门攻击评估。*IEEE Access* 7(2019)， 47230-47244。https://doi.org/10.1109/ACCESS.2019.2909068

[63] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. 2017。通过实体解析和加性同态加密在垂直分区数据上的私有联邦学习。CoRR abs/1711.10677(2017)。arXiv:1711.10677 http://arxiv.org/abs/1711.10677

[64] J. Henry Hinnefeld, Peter Cooman, Nat mamo, Rupert Deese。2018。评估存在数据集偏差的公平性指标。CoRR abs/1809.09245(2018)。arXiv:1809.09245 http://arxiv.org/abs/1809.09245

[65]刘建军，刘建军。2015。《在神经网络中提炼知识》。arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531

[66] Geoffrey E. Hinton, Simon Osindero, Yee Whye Teh. 2006。一种用于深度信念网络的快速学习算法。*神经计算* 18,7(2006)，1527-1554。DOI:https://doi.org/10.1162/neco.2006.18.7.1527

[67]胡胜山，尚兴灿，秦展，李明辉，王茜，王聪。2019。自动语音识别的对抗性示例:攻击和对策。*IEEE Communications Magazine* 57, 10(2019)， 120-126。

[68]胡玮炜，谭颖。2017。基于 RNN 的恶意软件检测算法的黑盒攻击。CoRR abs/1705.08131(2017)。arXiv:1705.08131 http://arxiv.org/abs/1705.08131

[69]胡玮玮，谭颖。2017。基于 GAN 的黑盒攻击对抗性恶意软件示例生成。CoRR abs/1702.05983(2017)。arXiv:1702.05983 http://arxiv.org/abs/1702.05983

[70] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019。对抗性的例子不是 bug，它们是特征。见《*第 33 届神经信息处理系统国际会议论文集*》，125-136。

[71] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018。操纵机器学习:回归学习的投毒攻击与对策。《*2018 IEEE 安全与隐私研讨会论文集*》，Vol. 2018- may。19-35。DOI:https://doi.org/10.1109/SP.2018.00057

[72]贾金元，龚振强，张扬，龚振强。Memguard:通过对抗性示例防御黑盒成员推理攻击。*2019 年 ACM SIGSAC 计算机与通信安全会议论文集*，259-274。

[73] Robin Jia和 Percy Liang. 2017。评价阅读理解系统的对抗性例子。*2017 年自然语言处理经验方法会议论文集*。计算语言学协会，斯特劳兹堡，2021-2031。

[74]金国庆，沈世伟，张东明，戴峰，张永东。2019。猿-氮化镓:对抗扰动消除与 GAN。*2019 年 IEEE 声学、语音和信号处理国际会议论文集*。IEEE 3842 - 3846。

[75] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, Mykel J. Kochenderfer. 2017。Reluplex:一种用于验证深度神经网络的高效 SMT 求解器。《*计算机辅助验证国际会议论文集*》。施普林格,97 - 117。

[76] Bedeuro Kim, Alsharif Abuadbba，高岩松，郑一峰，Muhammad Ejaz Ahmed, Hyoungshick Kim, Surya Nepal. 2020。Decamouflage:一种检测卷积神经网络图像缩放攻击的框架。CoRR abs/2010.03735(2020)。arXiv:2010.03735 https://arxiv.org/abs/2010.03735

[77]庞伟 Koh, Percy Liang. 2020。通过影响函数理解黑箱预测。arXiv:1703.04730 [stat.ML] https://arxiv.org/abs/1703.04730

[78]孔业豪，张纪良。2020。对抗性音频:一种新的信息隐藏方法。《*Interspeech 论文集*》，2287-2291。

[79] Denis Foo Kune, John Backes, Shane S. Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. 2013。鬼语:缓解针对模拟传感器的 EMI 信号注入攻击。*IEEE 安全与隐私研讨会论文集*，145-159。DOI:https://doi.org/10.1109/SP.2013.20

[80] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016。物理世界中的对抗性例子。CoRR abs/1607.02533(2016)。arXiv:1607.02533 http://arxiv.org/abs/1607.02533

[81]阿列克谢·库拉金，伊恩·j·古德费罗，萨米·本吉奥。2016。大规模的对抗性机器学习。《*第五届国际学习表征会议论文集*》。ICLR。

[82]李贤吉，韩成烨，李正宇。2017。生成对抗训练器:GAN 对抗摄动的防御。CoRR abs/1705.03387(2017)。arXiv:1705.03387 http://arxiv.org/abs/1705.03387

[83]李林阳，马若天，郭其鹏，薛向阳，邱希鹏。2020。BERT-Attack:利用 BERT 对 BERT 进行对抗性攻击。CoRR abs/2004.09984(2020)。arXiv:2004.09984 https://arxiv.org/abs/2004.09984

ACM 计算调查，第 55 卷，第 1 期，第 20 条。出版日期:2021 年 11 月。

[84]李文佳、Neha Bala、Aemun Ahmar、Fernanda Tovar、Arpit Battu、Prachi Bambarkar. 2019。针对对抗性示例攻击的android系统鲁棒恶意软件检测方法。*2019 年 IEEE 第五届协作与互联网计算国际会议论文集*。IEEE 360 - 365。

[85]李文佳，宋厚兵。2015。ART:一种保护车载自组织网络安全的抗攻击信任管理方案。*IEEE 智能交通系统学报*，17,4(2015)，960-969。

[86]李文佳，宋厚冰，曾峰。2017。智慧城市中基于策略的物联网安全可信感知。*IEEE 物联网杂志* 5,2(2017)，716-723。

[87]李鑫、李福新。2017。卷积滤波统计在深度网络中的对抗性样例检测。*IEEE 计算机视觉国际会议论文集*，5764-5772。

[88]梁伟，谢松友，蔡佳红，徐建波，胡玉鹏，徐阳，邱美康。2021。智能网络-物理系统中服务推荐的深度神经网络安全协同过滤方案。*IEEE 物联网杂志*(2021)，1-1。https://doi.org/10.1109/JIOT.2021.3086845

[89]林华清，闫铮，陈宇，张丽芳。2018。网络安全相关数据采集技术综述。*IEEE Access*6, 1(2018)， 18345-18365。

[90]刘畅，李波，Yevgeniy Vorobeychik, Alina Oprea, 2017。针对训练数据中毒的鲁棒线性回归。*第十届 ACM 人工智能与安全研讨会论文集*，91-102。DOI:https://doi。org/10.1145/3128572.3140447

[91]刘高，郑艳，Witold Pedrycz. 2018。移动 Ad Hoc 网络中攻击检测和安全测量的数据收集:综述。*网络与计算机应用学报* 105(2018)，105 - 122。https://doi。org/10.1016/j.jnca.2018.01.004

[92]刘康，布伦丹 Dolan-Gavitt . 2018。精细修剪:防范深度神经网络的后门攻击。*《攻击、入侵和防御研究国际研讨会论文集》*。施普林格,273 - 294。

[93]刘英奇、马世清、尤斯拉·阿弗尔、李文川、张祥宇。2017。神经网络上的特洛伊木马攻击。*《网络与分布式系统安全研讨会论文集》*。

[94] Auranuch Lorsakul 和 Jackrit Suthakorn. 2007。基于 OpenCV 的神经网络交通标志识别:走向智能车辆/驾驶员辅助系统。*《第四届泛在机器人与环境智能国际会议论文集》*，1-19。检索自 http://crit2007.bartlab.org/Dr.Jackrit’ papers /ney/1。交通{_}{_}Lorsakul迹象{_}ISR.pdf。

[95]吕家军，Theerasit Issaranon, David Forsyth. 2017。安全网:稳健地检测和拒绝对抗性示例。*IEEE 计算机视觉国际会议论文集*，446-454。

[96] Nitin Madnani 和 Bonnie J. Dorr. 2010。生成短语和句子释义:数据驱动方法综述。*计算语言学* 36,3(2010)，341-387。DOI:https://doi.org/10.1162/coli_a_00002

[97]约翰·麦卡锡。1956. 人工智能(AI)在达特茅斯创造。2021 年 10 月 28 日检索自 https://250。dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth。

[98]李晓明，李晓明，李晓明。布伦丹。通信——基于去中心化数据的深度网络高效学习。摘自*《人工智能与统计学论文集》*。PMLR, 1273 - 1282。

[99]梅世科、朱晓金。2015。利用机器教学识别对机器学习者的最优训练集攻击。摘自 *AAAI 人工智能会议论文集*。卷 29。

[100] Gys Albertus Marthinus Meiring and Hermanus Carel Myburgh. 2015。智能驾驶风格分析系统及相关人工智能算法综述。*传感器* 15,12 (Dec 2015)， 30653-30682。DOI:https://doi.org/10。3390 / s151229822

[101]孟冬雨、陈浩。2017。磁铁:对抗性例子的双管齐下防御。*2017 年 ACM SIGSAC 计算机与通信安全会议论文集*，135-147。

[102] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, Cynthia Rudin, 2020。PULSE:基于生成模型潜在空间探索的自监督照片上采样。*IEEE/CVF 计算机视觉与模式识别会议论文集*，2437-2445。

[103] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017 。关于检测对抗性扰动。arXiv:1702.04267 [stat.ML] https://arxiv.org/abs/1702.04267

[104] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017。普遍对抗性微扰。*《IEEE 计算机视觉与模式识别会议论文集》*，1765-1773。

[105] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016。Deepfool:一种简单而准确的愚弄深度神经网络的方法。*《IEEE 计算机视觉与模式识别会议论文集》*，2574-2582。

[106] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, Pascal Frossard, 2019。通过曲率正则化实现鲁棒性，反之亦然。*《IEEE 计算机视觉与模式识别会议论文集》*，9078-9086。

[107] Konda Reddy Mopuri, Aditya Ganeshan, R. Venkatesh Babu. 2018。用于制作通用对抗性微扰的可泛化无数据目标。*IEEE 模式分析与机器智能学报 41,10*(2018)，2452-2465。

[108] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi。2010。基于 Mel频率 Cep- stral系数(MFCC)和动态时间翘曲(DTW)技术的语音识别算法。CoRR abs/1003.4083(2010)。arXiv: 1003.4083 http://arxiv.org/abs/1003.4083

[109] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. 2017。走向用反向梯度优化毒化深度学习算法。*第十届ACM 人工智能与安全研讨会论文集*，27-38。

[110] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. 2008。利用机器学习颠覆你的垃圾邮件过滤器。在*第一届大规模利用和紧急威胁 USENIX 研讨会论文集:僵尸网络，间谍软件，蠕虫等*。

[111]亚历山德拉·奥尔特亚努，卡洛斯·卡斯蒂略，费尔南多·迪亚兹和埃姆雷 Kı cı man。2019. 社会数据:偏见、方法论陷阱和伦理界限。*《大数据前沿 2》*(2019)，13。https://doi.org/10.3389/fdata.2019.00013

[112]道格拉斯·奥肖内西。2008. 自动语音识别:历史、方法和挑战。*模式识别* 41,10(2008)，2965-2979。

[113] Mesut Ozdag. 2018。对深度神经网络的对抗性攻击和防御:综述。*《计算机科学进展》* 140(2018)，152-161。https://doi.org/10.1016/j.procs.2018.10.315Cyber Physical Systems and Deep Learning芝加哥，伊利诺斯州，2018年 11月 5-7日。

[114] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016。对抗性环境下深度学习的局限性。*2016 年IEEE 欧洲安全与隐私研讨会论文集*。DOI:https://doi.org/10.1109/EuroSP.2016.36

蒸馏作为对抗深度神经网络对抗性扰动的防御。*2016 年 IEEE 安全与隐私研讨会论文集*。IEEE 582 - 597。

[116] George Philipp, Jaime G. Carbonell. 2018。非线性系数-预测深度神经网络中的过拟合。*CoRR abs/1806.00179(2018)*。arXiv:1806.00179 http://arxiv.org/abs/1806.00179

[117] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, peter Motlicek，钱彦敏，peter Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely。2011。The Kaldi 语音识别工具箱。*《IEEE 2011 年自动语音识别与理解研讨会论文集》*。IEEE信号处理学会。

[118] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, James stover。2018。用像素偏转来偏转对抗性攻击。*《IEEE 计算机视觉与模式识别会议论文集》*，8571 - 8580。

[119] Katyanna Quach. 2020。研究人员制作了一个 OpenAI GPT-3医疗聊天机器人作为实验。它告诉一个模拟病人自杀。2021年 10月 28日检索自 https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_ experiment/。

[120] Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. 2020。对抗性预处理:理解和防止机器学习中的图像缩放攻击。第 29 *届 USENIX 安全研讨会论文集*，1363-1380。

[121] Erwin Quiring and Konrad Rieck。2020。基于图像缩放攻击的后门和毒化神经网络。*CoRR abs/2003.08633(2020)*。arXiv:2003.08633 https://arxiv.org/abs/2003.08633

[122] Aditi Raghunathan, Jacob Steinhardt, Percy Liang. 2018。对抗性实例的认证防御。*CoRR abs/1801.09344(2018)*。arXiv:1801.09344 http://arxiv.org/abs/1801.09344

[123] Kanishka Rao, ha<e:1> im Sak, Rohit Prabhavalkar. 2017。利用 rnn-换能器探索流端到端语音识别的架构、数据和单元。*2017 年 IEEE 自动语音识别与理解研讨会论文集*，193-199。

[124] Konda Reddy Mopuri, Phani Krishna Uppala, and R. Venkatesh Babu. 2018。询问、获取和攻击:使用类别印象生成无数据的 UAP。*《欧洲计算机视觉会议论文集》*，19-34。

[125] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, R. Venkatesh Babu, 2018。NAG:用于对手生成的网络。*《IEEE 计算机视觉与模式识别会议论文集》*，742-751。

[126] Mohsen Rezvani, Aleksandar Ignjatovic, Elisa Bertino, and Sanjay Jha. 2014。存在合谋攻击的无线传感器网络安全数据聚合技术。*IEEE 可靠与安全计算学报* 12,1(2014)，98-110。

[127]黄玲，刘成汉，刘志强，刘志强，刘志强。解毒剂:理解和防御异常探测器的中毒。*ACM SIGCOMM 互联网测量会议论文集*。DOI:https://doi.org/10.1145/1644893.1644895

[128] Pouya Samangouei, Maya Kabkab, Rama Chellappa. 2018。 Defense-GAN:使用生成模型保护分类器免受对抗性攻击。 *CoRR abs/1805.06605(2018)*。arXiv:1805.06605 http://arxiv.org/abs/1805.06605

[129]佐藤元树、铃木俊、新藤博之、松本裕司。2018。文本输入嵌入空间中的可解释对抗摄动。*CoRR abs/1805.02917(2018)*。arXiv:1805.02917 http://arxiv.org/abs/1805.02917

[130]刘建军，黄晓明，张晓明，张晓明。毒青蛙!针对神经网络的清洁标签投毒攻击。*第 32 届神经信息处理系统国际会议论文集*。6103-6113。

[131]孙允木，孙永明，沈永杰，虎哲，朴永锡，金永大。2016。采样竞赛:在模数系统上绕过基于时序的模拟有源传感器欺骗检测。*第十届 USENIX 进攻技术研讨会论文集*。

[132] Yasser Shoukry, Paul Martin, Paulo Tabuada, and Mani Srivastava. 2013。针对防抱死制动系统的非侵入式欺骗攻击。在 *加密硬件和嵌入式系统*。G. Bertoni 和 J. S. Coron(主编)，《计算机科学课堂笔记》(包括人工智能和生物信息学的子系列课堂笔记)，第 8086 卷，Springer，55-72 页。DOI:https://doi.org/10.1007/978-3-642-40349-1-4

[133] Yasser Shoukry, Paul Martin, Yair Yona, Suhas Diggavi, and Mani Srivastava. 2015。PyCRA:欺骗攻击类别和主题描述符下主动传感器的物理挑战-响应认证。*第 22 届 ACM SIGSAC 计算机与通信安全会议论文集*，1004-1015。

[134] Marco della Cava. 2018。Uber 自动驾驶汽车杀死亚利桑那州的行人，实现了对新技术的最大恐惧。检索自 2021 年 10 月 28 日 https://www.usatoday.com/story/tech/2018/03/19/uber-self-driving-car-kills-arizona- woman/438473002/。

[135]孙允木，虎哲，金永大，朴永锡，卢柱焕，金永大。2015。陀螺仪传感器上带有有意声噪声的摇摆无人机。第 24 *届 USENIX 安全研讨会论文集*。

[136]宋从正和 Vitaly Shmatikov. 2018。用对抗性文本图像欺骗 OCR 系统。*CoRR abs/1802.05385(2018)*。arXiv:1802.05385 http://arxiv.org/abs/1802.05385

[137]宋利伟，Reza Shokri, Prateek Mittal. 2019。针对对抗性示例保护机器学习模型的隐私风险。*2019 年 ACM SIGSAC 计算机与通信安全会议论文集*，241-257。

[138]宋洋，金永大，王晓明，王晓明。2017。PixelDefend:利用生成模型来理解和防御对抗性示例。*CoRR abs/1710.10766(2017)*。arXiv:1710.10766 http://arxiv.org/abs/1710.10766

[139] Jacob Steinhardt, Pang Wei Koh, Percy Liang. 2017。数据投毒攻击的认证防御。*第 31 届神经信息处理系统国际会议论文集*。Vol. 2017-Decem，3518-3530。

[140] Carsten Stephan, Henning Cammann, Axel Semjonow, Eleftherios P. Diamandis, Leon F. A. Wymenga, Michael Lein, Pranav Sinha, Stefan A. Loening, and Klaus Jung. 2002。多中心评价人工神经网络提高前列腺癌检出率，减少不必要的活检。*临床化学* 48,8(2002)，1279 - 1287。DOI:https://doi.org/10.1093/clinchem/48.8.1279

[141]苏佳伟、达尼洛·瓦斯孔塞洛斯·瓦尔加斯、樱井 Kouichi . 2019。用于愚弄深度神经网络的一个像素攻击。*IEEE 进化计算学报* 23,5(2019)，828-841。

[142]孙立超，王骥，Philip S. Yu，李波。2018。图数据的对抗性攻防:综述。*CoRR abs/1812.10528(2018)*。arXiv:1812.10528 http://arxiv.org/abs/1812.10528

[143]孙梦颖，唐凤毅，易金峰，王飞，周嘉宇。2018。通过对深度预测模型的对抗性攻击，识别医疗记录中的易感位置。在 *ACM SIGKDD 知识发现与数据挖掘国际会议论文集*上。DOI:https://doi.org/10.1145/3219819.3219909

[144]孙思宁，叶青峰，Mari Ostendorf，黄美玉，谢磊。2018。鲁棒语音识别的对抗性示例训练增强。*CoRR abs/1806.02782(2018)*。arXiv:1806.02782 http://arxiv.org/ abs/1806.02782

[145] Latanya Sweeney. 2000。简单的人口统计数据往往能独特地识别人。*健康*，2000(2000)，1-34。

[146]刘建军，刘建军，刘建军，刘建军。神经网络的耐人寻味性质。*《第二届国际学习表征会议论文集》*，1-10。

[147] Yasmin Tadjdeh. 2017。DARPA 的“AI next”项目开花结果。*NDIA 的《商业与技术》杂志*。检索自 https://www.nationaldefensemagazine.org/articles/2019/7/2/algorithmic-warfare-darpas-ai- next-program-bearing-fruit。

[148]张晓明，张晓明，张晓明，等。2008。黑箱音频系统的目标对抗性示例。*CoRR abs/1805.07820(2018)*。arXiv:1805.07820 http://arxiv.org/abs/1805.07820

[149]弗洛里安·特拉米特，阿列克谢·库拉金，尼古拉斯·派珀诺特，伊恩·j·古德费罗，丹·博内，帕特里克·d·麦克丹尼尔。2018.综合对抗训练:攻击和防御。第*六届国际学习表征会议论文集，ICLR 2018，温哥华，BC，加拿大，2018年4月30日-5月3日*。OpenReview.net。https://openreview.net/论坛？id = rkZvSe-RZ。

[150] Florian tram<e:1>，张帆，Ari Juels, Michael K. Reiter, Thomas Ristenpart. 2016。通过预测api窃取机器学习模型。第25*届USENIX安全研讨会论文集*，601-618。

[151] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019。标签一致后门攻击。arXiv:1912.02771 [stat.ML] https://arxiv.org/abs/1912.02771

[152] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni. 2019。恶意软件分析的机器学习技术综述。*计算机与安全*81(2019)，123-147。https://doi.org/10.1016/j.cose.2018.11.001

[153] Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz J. Och, and Juri Ganitkevitch. 2011。结构化预测输出的水印及其在统计机器翻译中的应用。*《自然语言处理中的经验方法会议论文集》*。计算语言学协会，1363-1372。

[154]詹姆斯·文森特。2020。一个让奥巴马变白的机器学习工具能(也不能)告诉我们关于AI偏见的什么。2021年10月28日检索自 https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-toolpulse-stylegan-obama-bias。

[155] Putra Wanda, Huang Jin Jie. 2020。DeepProfile:利用动态CNN在在线社交网络中发现虚假个人资料。*信息安全与应用学报*52(2020)，102465。https://doi.org/10.1016/j.jisa.2020.102465

[156]王博伦，姚元顺，单肖恩，李慧英，毕姆·维斯瓦纳特，郑海涛，赵本毅。2019。神经清洗:识别和缓解神经网络中的后门攻击。*2019年IEEE安全与隐私研讨会论文集*。IEEE 707-723。

[157]王志波，宋孟凯，张志飞，宋洋，王茜，齐海荣。2019。超越推断类代表——代表:来自联邦学习的用户级隐私泄露。*《IEEE计算机通信会议论文集》*。IEEE 2512-2520。

[158]文冠超，胡宇鹏，陈江，曹娜，郑勤。2017。一种基于图像纹理和BP神经网络的云存储系统恶意文件检测技术。*2017年IEEE计算机通信研讨会论文集*。IEEE 426-431。

[159]夏飞，刘瑞山。2016。基于生成式对抗网络的对抗例生成与防御。*arXiv预印本arXiv:1712.00170*(2016)。

[160]黄晓、巴蒂斯塔·比吉奥、加文·布朗、乔治·富梅拉、克劳迪娅·埃克特、法比奥·罗利。2015。特征选择对训练数据中毒安全吗?*机器学习国际会议论文集*，1689-1698。

[161]肖启学，陈玉飞，沈超，陈宇，李康。眼见为实:伪装攻击图像缩放算法。*《第28届USENIX安全研讨会论文集》*，443-460页。

[162]肖启学，李康，张德月，徐伟林。2018。深度学习实现中的安全风险。*2018年IEEE安全与隐私研讨会论文集*。IEEE 123-128。

[163]谢慈航，王建宇，张志帅，任周，Alan Yuille. 2017。通过随机化减轻对抗性效应。*CoRR abs/1711.01991(2017)*。arXiv:1711.01991 http://arxiv.org/abs/1711.01991

[164]徐桂标，曹政，胡宝刚，何塞·c·普林西比。2017。基于重标度铰链损失函数的鲁棒支持向量机。*模式识别*100,63(2017)，139-148。DOI:https://doi.org/10.1016/j.patcog.2016。09.045

[165]徐涵、马瑶、刘浩晨、Debayan Deb、刘辉、唐纪良、Anil K. Jain, 2020。图像、图形和文本中的对抗性攻击与防御:综述。*国际自动化与计算杂志*17,2(2020)，151-178。

[166]徐伟林，David Evans，齐彦军。2017。特征压缩:深度神经网络中对抗性样本的检测。*CoRR abs/1704.01155(2017)*。arXiv:1704.01155 http://arxiv.org/abs/1704.01155

[167]徐伟林，齐彦军，David Evans. 2016。自动回避分类器。*2016网络与分布式系统研讨会论文集*。卷。10。

[168]张志强，张志强。2018。物理攻击的鲁棒音频对抗性示例。*CoRR abs/1810.11793(2018)*。arXiv:1810.11793 http://arxiv.org/abs/1810.11793

[169]杨朝飞，吴清，李海，陈怡然。2017。针对神经网络的生成投毒攻击方法。*CoRR abs/1703.01340(2017)*。arXiv:1703.01340 http://arxiv.org/abs/1703.01340。

[170]叶艳芳，李涛，Donald Adjeroh, S. Sitharama Iyengar. 2017。基于数据挖掘技术的恶意软件检测研究综述。*ACM计算调查*50,3(2017)，1-40。

[171]袁学静，陈宇轩，赵悦，龙云辉，刘康康，陈凯，张生智，黄鹤清，王晓峰，Carl A. Gunter。2018。Commandersong:一种实用对抗性语音识别的系统方法。第27*届USENIX安全研讨会论文集*，49-64。

[172]张晨月，李文嘉，罗元生，胡宇鹏。2020。AIT:基于区块链技术的基于 ai的车联网信任管理系统。*IEEE 物联网学报*，8,5(2020)，3157-3169。

[173]张国明，闫晨，纪晓宇，张天辰，张太民，徐文渊。2017。海豚攻击:听不清的语音指令。*2017 ACM SIGSAC 计算机与通信安全会议论文集*，103-117。

[174]张纪良、李晨。2020。对抗性实例:机遇与挑战。*IEEE 神经网络与学习系统学报 31,7*(2020)，2578-2593。DOI:https://doi.org/10.1109/TNNLS.2019.2933524

[175]张纪良、李晨、叶静、曲刚。2020。机器学习中的隐私威胁与保护。*《2020 年 VLSI 大湖研讨会论文集》*。531 - 536。

[176]张继良，彭爽，胡宇鹏，彭飞，胡宇鹏，赖金梅，叶静，王祥奇。HRAE:对抗对抗性示例攻击的硬件辅助随机化。*在 2020 年 IEEE 第 29 届亚洲测试研讨会论文集中*。IEEE，1 - 6。

2021 年 2 月收稿;2021 年 9 月修订;2021 年 9 月接受