计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190437
56(10):2049-2070,2019

机器学习系统的隐私和安全问题综述

何英哲　胡兴波　何锦雯　孟国柱　陈　恺
(信息安全国家重点实验室(中国科学院信息工程研究所)　北京　100195)
(中国科学院信息工程研究所　北京　100195)
(中国科学院大学网络空间安全学院　北京　101408)
(heyingzhe@iie.ac.cn)

Privacy and Security Issues in Machine Learning Systems: A Survey

He Yingzhe, Hu Xingbo, He Jinwen, Meng Guozhu, and Chen Kai
(*State Key Laboratory of Information Security* (*Institute of Information Engineering*, *Chinese Academy of Sciences*), *Beijing* 100195)
(*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100195)
(*School of Cyber Security*, *University of Chinese Academy of Sciences*, *Beijing* 101408)

Abstract Artificial intelligence has penetrated into every corners of our life and brought humans great convenience. Especially in recent years, with the vigorous development of the deep learning branch in machine learning, there are more and more related applications in our life. Unfortunately, machine learning systems are suffering from many security hazards. Even worse, the popularity of machine learning systems further magnifies these hazards. In order to unveil these security hazards and assist in implementing a robust machine learning system, we conduct a comprehensive investigation of the mainstream deep learning systems. In the beginning of the study, we devise an analytical model for dissecting deep learning systems, and define our survey scope. Our surveyed deep learning systems span across four fields-image classification, audio speech recognition, malware detection, and natural language processing. We distill four types of security hazards and manifest them in multiple dimensions such as complexity, attack success rate, and damage. Furthermore, we survey defensive techniques for deep learning systems as well as their characteristics. Finally, through the observation of these systems, we propose the practical proposals of constructing robust deep learning system.

Key words machine learning security; deep learning security; attack and defense race; adversarial attack; membership inference attack; privacy-preserving

摘　要 人工智能已经渗透到生活的各个角落,给人类带来了极大的便利.尤其是近年来,随着机器学习中深度学习这一分支的蓬勃发展,生活中的相关应用越来越多.不幸的是,机器学习系统也面临着许多

收稿日期:2019-06-11;**修回日期**:2019-08-21
基金项目:国家重点研发计划项目(2016QY04W0805);国家自然科学基金项目(U1836211, 61728209);中国科学院青年创新促进会;北京市科技新星计划;北京市自然科学基金项目(JQ18011);国家前沿科技创新项目(YJKYYQ20170070)
This work was supported by the National Key Rearch and Development Program of China (2016QY04W0805), the National Natural Science Foundation of China (U1836211, 61728209), the Program of Youth Innovation Promotion Association CAS, the Beijing Nova Program, the Beijing Natural Science Foundation (JQ18011), and the National Frontier Science and Technology Innovation Project (YJKYYQ20170070).
通信作者:孟国柱(mengguozhu@iie.ac.cn)

安全隐患,而机器学习系统的普及更进一步放大了这些风险.为了揭示这些安全隐患并实现一个强大的机器学习系统,对主流的深度学习系统进行了调查.首先设计了一个剖析深度学习系统的分析模型,并界定了调查范围.调查的深度学习系统跨越了 4 个领域——图像分类、音频语音识别、恶意软件检测和自然语言处理,提取了对应 4 种类型的安全隐患,并从复杂性、攻击成功率和破坏等多个维度对其进行了表征和度量.随后,调研了针对深度学习系统的防御技术及其特点.最后通过对这些系统的观察,提出了构建健壮的深度学习系统的建议.

关键词 机器学习安全;深度学习安全;攻防竞赛;对抗攻击;成员推理攻击;隐私保护

中图法分类号 TP391

深度学习的广泛应用所带来的成功并不能保证其安全性,新的威胁和攻击每天都在出现,它们危及深度学习模型,进而危及人们的隐私、金融资产和安全.作为一种新兴的技术,深度学习的安全问题往往被忽视.因此,系统地研究深度学习的安全问题并进一步提出有效的措施,是迫切而关键的.

深度学习已广泛应用于图像分类、语音识别、自然语言处理、恶意软件检测等多个领域.由于计算能力的巨大进步和数据量的急剧增加,深度学习在这些场景中显示出了优越的潜力.深度学习尤其擅长无监督特征学习,加深对一个对象的理解,具有强大的预测能力.然而,深度学习正遭受精心策划的攻击所带来的一系列威胁.例如深度学习系统很容易被对抗样本所欺骗,从而导致错误的分类.另一方面,使用在线深度学习系统进行分类的用户不得不向服务器公开他们的数据,这会导致隐私泄露.更糟糕的是,深度学习的广泛使用加剧了这些安全风险.

研究人员正在探索和研究针对深度学习系统的潜在攻击以及相应的防御技术.文献[1]是探索神经网络安全性的先驱,Szegedy 等人用难以察觉的扰动(对抗样本)揭示了神经网络的脆弱特性.自此以后,对抗攻击迅速成为人工智能和安全领域的热门术语.许多工作都致力于披露不同深度学习模型(例如深度神经网络(DNN)、卷积神经网络(CNN)、循环神经网络(RNN))中的漏洞和提高对抗样本的健壮性[2].另一方面,深度学习系统的大量商业部署提出了对专有资产(如训练数据[3-6]、模型参数[7-10])保护的要求,它引发了一场“军备竞赛”,在这场竞争中,攻击者从竞争对手那里偷取隐私信息,而相应的防御者则采取广泛的措施来抵御攻击.

为了全面了解深度学习中的隐私和安全问题,我们对相关文献和系统进行了调查,研究了 150 篇左右的相关研究,跨越了图像分类、语音识别、自然语言处理和恶意软件检测 4 个领域.由于很难完成包罗万象的调查,所以我们选择了更具代表性的研究:例如那些在著名会议和期刊上获得发表的研究;虽然只发表在研讨会或专题讨论会上,但被引用次数高(超过 50 次)的研究;在公共平台上(如 arXiv)最近发表的热点方向论文.基于调研工作,我们将这些攻击归纳为 4 类:模型提取攻击(model extraction attack)、模型逆向攻击(model inversion attack)、投毒攻击(poisoning attack)和对抗攻击(adversarial attack).其中,模型提取和逆向攻击针对的是隐私,前者主要窃取模型的信息,后者主要获得训练数据集的信息;投毒攻击和对抗攻击针对的是安全,前者主要在训练阶段投放恶意数据从而降低模型的分类准确率,后者主要在预测阶段制造对抗样本来欺骗模型.

图 1 展示了过去 5 年与机器学习系统安全有关的研究数据,包括对模型的各种攻击以及隐私保护、安全防御等研究.在过去的 5 年里,相关研究的数量急剧增长,2017 年增长 100%,2018 年增长 61.5%,近 2 年的文章数量占了接近 70%,这也说明了深度学习、机器学习乃至人工智能领域的安全问题越来越引起人们的重视.

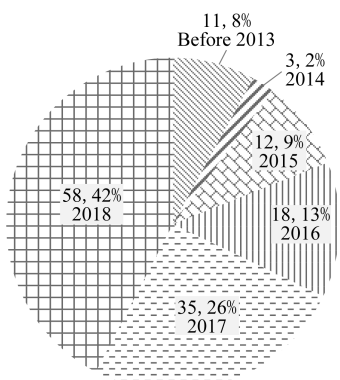

Fig. 1 The number of publications in recent years

图 1 近年来相关研究数量

图 2 显示了我们所研究的 4 类攻击的相关研究数量.其中对抗攻击是最引人注目的,对模型实施对抗攻击的研究占据了 50%,它可以直接使模型判断错误,因此威胁范围很广.模型提取攻击作为近年来新兴的攻击类型,由于其奠基性(模型提取攻击获得的模型可以为其他攻击提供白盒基础),难度较大,故相关的研究数量最少,未来还有很大的研究空间.我们调研的文章主要来自人工智能社区和安全社区,其中大部分来自人工智能社区.根据发表地点来对二者区分,具体来说,ICML,CVPR,AAAI,IJCAI,TPAMI 等属于人工智能社区,IEEE S&P,CCS,USENIX Security,NDSS,AISec 等属于安全社区.

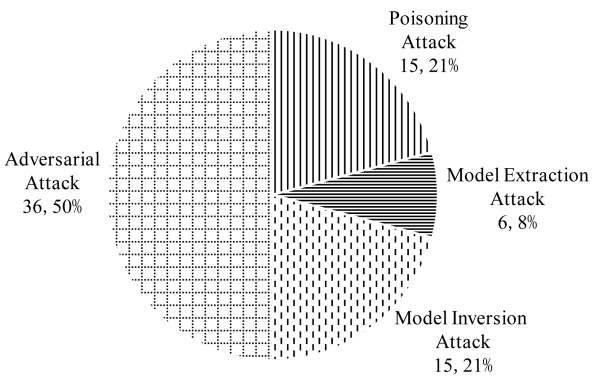

Fig. 2 Numbers of related researches on different attack types

图 2 不同攻击类型的相关研究数量

本文主要研究机器学习安全的范围、整个学习系统基本的组成部分、攻击方法、防御措施、实用性评价以及有价值的现象与结论,主要包含 4 方面贡献:

1) 攻击和防御技术的系统分析.总结了 4 种攻击类型和 3 种防御类型,全面地对机器学习系统的隐私和安全问题进行了调研和总结.

2) 机器学习系统的模块化.对机器学习系统进行剖析,按准备过程、训练过程、预测过程的时间线,按训练数据集、训练算法、模型结构、模型参数、预测数据及结果的空间线,系统地总结了机器学习的安全知识.

3) 各个攻击和防御类型内部具体技术的划分.对每种攻击防御类型内的攻防技术进行了剖析,将庞杂的技术文章进行分类,并分析了不同技术之间的差异和优劣.

4) 通过对机器学习系统的安全问题的观察和总结,以及对这些攻击和防御技术的分析和研究,提出了构建安全健壮的机器学习系统和保护机器学习所有参与者隐私安全的经验和建议.

1 相关工作

目前已有部分文献对机器学习的攻击和防御进行了调研和评估.在早期的工作中,Barreno 等人[11]对机器学习安全性进行了调研,并对针对机器学习系统的攻击进行了分类.他们在一个统计垃圾邮件的过滤器上进行了实验,从攻击的操作方式、对输入的影响和普遍性 3 个维度对攻击进行了剖析.Amodei 等人[12]介绍了机器学习中与事故风险相关的 5 个可能的研究问题,并根据其工作原理,以清洁机器人为例,讨论了可能的解决方法.

Papernot 等人[13]回顾了之前关于机器学习系统攻击和相应防御的工作.与以往的调研和综述不同,他们针对的是关于安全威胁的全面文献综述.Bae 等人[14]总结了安全与隐私概念下关于 AI 的攻击与防御方法.他们在黑盒子和白盒子里检查对抗和投毒攻击.随后,Papernot 等人[15]系统地研究了机器学习的安全性和隐私性,并提出了一种机器学习的威胁模型.他们按照训练过程和预测过程、黑盒模型和白盒模型的分类来介绍攻击方法.但他们没有过多涉及应用广泛的深度学习模型.Liu 等人[16]主要关注机器学习的 2 个阶段,即训练阶段和预测阶段,并提供了较全面的文献综述.他们将相应的防御措施分为 4 类.另外,他们的研究更关注对抗样本导致的数据分布漂移和机器学习算法导致的敏感信息泄露等问题.

Akhtar 等人[17]全面研究了计算机视觉领域中深度学习受到的对抗攻击,总结了 12 种不同类别的攻击方法.除常用的 CNN 外,他们还研究了对其他模型的攻击(如自动编码器、生成模型、RNN)以及物理世界中的攻击,此外他们也总结了多种防御方法.然而,这项工作的研究内容只限于计算机视觉领域的对抗攻击.Ling 等人开发的 DeepSec[18]是一个统一的评测平台.DeepSec 集成了对抗学习中 16 种攻击方法和 13 种防御方法,旨在衡量深度学习模型的脆弱性,并评估各种攻击和防御的有效性.

本文对机器学习系统特别是深度学习中的隐私和安全问题进行调研和总结,对攻击和防御方法进行分类,分析不同类别下的攻防技术,并介绍其在图像分类、语音识别、自然语言处理和恶意软件检测等不同领域的应用.

2 机器学习概述

2.1 机器学习系统

有监督的机器学习主要分为 2 个阶段:模型训练阶段和模型预测(推理)阶段.模型训练阶段将训练数据集作为输入,最后生成模型;模型预测阶段接受用户或攻击者的输入并提供预测结果.为了完成这 2 个阶段,模型设计人员必须指定使用的训练数据和训练算法.模型训练阶段生成经过调优的训练模型以及相关参数.而在运行训练算法之前,传统机器学习需要人工提取和选择特征,深度学习则委托训练算法自动识别可靠而有效的特征.通常,经过训练的模型可以部署用于商业用途.在商业应用中,模型根据接收到的输入计算最可能的结果.以恶意软件检测为例,安全分析人员首先从恶意软件中收集数据(可能是原始数据),提取有代表性的特征并构建分类模型,以检测恶意软件.

深度学习是机器学习这个广泛的家族的一部分,深度神经网络受到生物神经系统的启发,由成千上万个神经元组成,用来传递信息.深度学习受益于人工神经网络,通常使用更多的层来提取和转换特征.

为了使机器学习系统的过程形式化,我们在表 1 中给出了一些符号.给定一个机器学习任务,收集的数据可以表示为 $\boldsymbol{x}=(x^{(1)},x^{(2)},\cdots,x^{(n)})$.数据集 D 就是很多 $\boldsymbol{x}$ 组成的集合.假设 F 是一个机器学习系统,它可以根据给定的输入 $\boldsymbol{x}$,计算相应的结果 $\boldsymbol{y}$,即 $\boldsymbol{y}=F(\boldsymbol{x})$.在模型训练过程中,使用损失函数测量对真实结果的预测误差,训练过程希望通过微调参数获得最小的误差值.损失函数可以计算为 $L=\sum_{1\leqslant i\leqslant n}\|y_p^{(i)}-F(x^{(i)})\|^2$,其中 $\boldsymbol{y}_p$ 表示真实结果.因此模型训练过程可以表示为

$$\arg\min_F L.$$

Table 1 Formalization in Machine Learning System

表 1 机器学习系统的符号化

Symbol	Definition
D	Dataset
$x^{(1)},x^{(2)},\cdots,x^{(n)}$	Input Data
$y^{(1)},y^{(2)},\cdots,y^{(n)}$	Output Result
F	Model
w_{ij}^k	Weights Parameters
b_j^k	Bias Parameters
λ	Hyperparameters
$\boldsymbol{x}_t$	Prediction Input
$\boldsymbol{y}_t$	Prediction Output
δ	Perturbation

2.2 安全威胁

图 3 展示了一个经典的深度学习模型在训练阶段、预测阶段的过程容易受到的威胁.最近的研究表明,机器学习系统是脆弱的,很容易受到特定攻击的影响.根据攻击目标,这些攻击可以分为 4 类:投毒

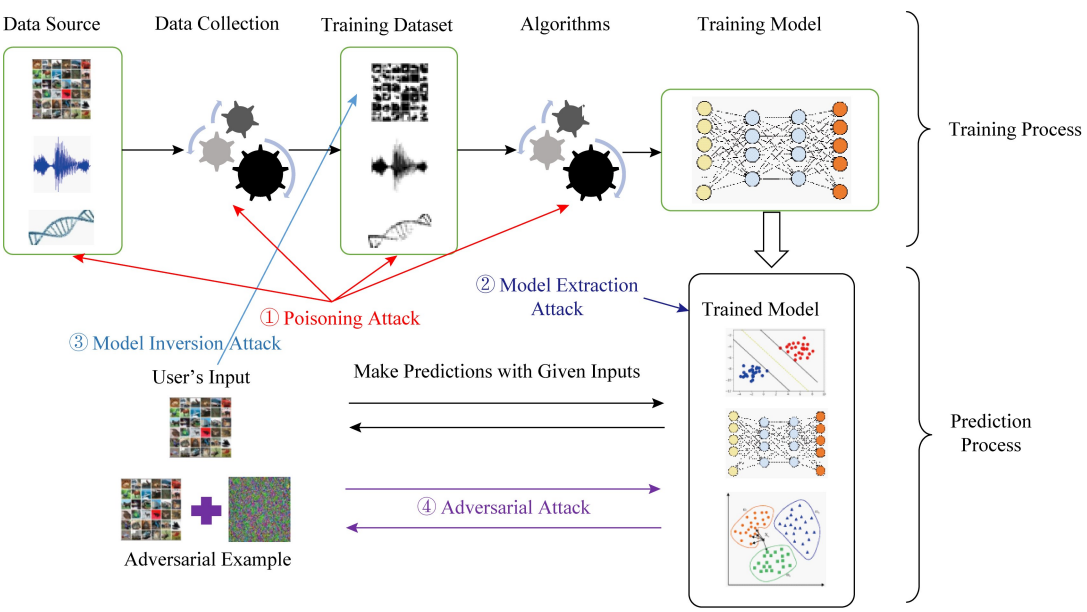

Fig. 3 Overview of attacks in machine learning system

图 3 机器学习系统攻击概述

攻击、模型提取攻击、模型逆向攻击和对抗攻击.在本节中,我们将通过示例及其正式定义来详细说明这些攻击.

1) 投毒攻击.投毒攻击主要是指在训练或再训练过程中,通过攻击训练数据集或算法来操纵机器学习模型的预测.由于在安全机器学习领域中,数据通常是非平稳的,其分布可能随时间而变化,因此一些模型不仅在训练过程中生成,而且在周期性再训练过程中随时间而变化.攻击训练数据集的方法主要包括污染源数据、向训练数据集中添加恶意样本、修改训练数据集中的部分标签、删除训练数据集中的一些原有样本等.攻击算法利用了不安全的特征选择方法或训练过程算法的弱点.投毒攻击会增加训练合适模型的难度.它还可以在生成的模型中为攻击者添加一个后门,攻击者可以使模型的预测偏向他想要的方向[19].

2) 模型提取攻击.模型提取攻击发生在训练好的模型上,主要用于窃取模型参数及非法获取模型.它违反了训练模型的保密性.在新的业务机器学习即服务(machine learning as a service, MLaaS)设置中,模型本身托管在一个安全的云服务中,它允许用户通过基于云的预测 API 查询模型.模型所有者通过让用户为预测 API 付费来实现模型的业务价值,所以机器学习模型属于商业秘密.此外,一个模型的训练过程需要收集大量的数据集,也需要大量的时间和巨大的计算能力,所以一旦提取出模型并对其滥用,就会给模型拥有者带来巨大的经济损失.

3) 模型逆向攻击.在早期的认识中,训练数据集和训练模型之间只有一个信息流,即从数据集到模型.事实上,许多研究表明还存在一个逆向信息流,即从模型信息中恢复数据集信息,这称为模型逆向攻击.模型逆向攻击是指将训练数据集信息从模型中逆向提取出来.它主要包括成员推理攻击(membership inference attack, MIA)和属性推理攻击(property inference attack, PIA).MIA 主要对数据集中是否出现特定记录进行推断,即判断隶属度,这是目前研究的热点.PIA 则主要获取数据集的如性别分布、年龄分布、收入分布、患病率等属性信息.模型逆向攻击窃取了训练数据集中成员的私有信息,也损害了数据集所有者的商业价值.发生这种情况有 2 个原因:①不充分的隐私保护,如信息泄露[20];②不安全的算法[21].为了加强对个人隐私的保护,欧盟于 2018 年颁布 GDPR,它明确界定了个人资料的隐私,并对其进行严格保护[22].

4) 对抗攻击.对抗攻击是指将对抗样例提交到训练好的模型中,从而使模型预测错误,它也被称为逃避攻击(evasion attack).对抗样本是从原来正常的样本上添加了轻微的扰动,可以导致分类模型分类错误的样本.对抗样本另外一个特点是仅造成模型分类错误,人还是可以将它正确分类.同样,在语音和文本识别领域,对抗样本也未对原文进行令人察觉的修改.在恶意软件检测领域,恶意软件作者在其软件上添加一些特殊的语句可以逃避反病毒软件的检测.

3 隐　　私

隐私是信息安全领域一个普遍存在但又难以解决的问题.广义上说,隐私包括有价值的资产和数据不受窃取、推断和干预的权利.由于深度学习是建立在海量数据之上的,经过训练的模型实际上是一个数据模型,而经过训练的模型需要与来自个人的测试数据进行大量交互,因此隐私显得更加重要,也需要更强的保护.在本节中,我们将介绍深度学习系统中存在的隐私问题,并从攻击和防御 2 个方面介绍当前的研究现状.

3.1 隐私问题简介

从本质上讲,深度学习将大量的数据转换为一个数据模型,该数据模型可以进一步地根据输入数据预测结果,凡是涉及到数据的部分都需要关注其隐私问题.基于整个深度学习过程,我们将隐私保护的对象分类为:1)训练数据集;2)模型结构、算法和模型参数;3)预测数据与结果.

高质量的训练数据对深度学习的表现至关重要.一般来说,训练数据的收集是一个耗时耗钱的过程:来自互联网的免费数据集通常不符合要求;从专业公司购买数据需要花费大量金钱;手工标记数据需要花费很多时间.此外,训练数据在最终传递到深度学习系统之前,还需要经过清洗、去噪和过滤等过程.因此,训练数据对于一个公司来说是至关重要的,也是非常有价值的,它的泄露意味着公司资产的损失.

深度学习中的训练模型是一种数据模型,是训练数据的抽象表示.在现代深度学习系统中,训练阶段需要处理大量的数据和多层训练,对高性能计算和海量存储有着严格的要求.也就是说,经过训练的模型被认为是深度学习系统的核心竞争力.通常,训练模型包含 3 种类型的数据资产:1)模型,例如传统

的机器学习和深度神经网络;2)超参数,设计了训练算法的结构如网络层数和神经元个数;3)参数,为多层神经网络中一层到另一层的计算系数.

在这种情况下,经过训练的模型具有极其重要的商业和创新价值.一旦模型被复制、泄露或提取,模型所有者的利益将受到严重损害.在预测输入和预测结果方面,隐私来自于深度学习系统的使用者和提供者.恶意的服务提供者可能会保留用户的预测数据和结果,以便从中提取敏感信息,或者用于其他目的.另一方面,预测输入和结果可能会受到不法分子的攻击,他们可以利用这些数据来为自己创造利润.

3.2 隐私问题研究工作

为了对隐私问题提供一个全面的概述,我们调查了48篇相关的文章,21篇与破坏隐私相关的文章和27篇与保护隐私相关的文章.

目前主流的隐私破坏方法主要有模型提取攻击(model extraction attack)和模型逆向攻击(model inversion attack).二者的主要区别是,前者关注模型的隐私信息,后者关注数据集的隐私信息.在模型提取攻击中,攻击者通过深度学习系统提供的API向模型发送大量的预测数据,然后接收模型返回的类标签和置信度系数,计算出模型的参数,最后还原原始模型.这种攻击可以破坏模型本身的隐私,损害模型所有者的利益,为攻击者创造商业价值,还可以帮助实现模型逆向攻击和对抗攻击.

在模型逆向攻击中,攻击者通过向模型提供预测数据得到模型的置信度系数,破坏用户或数据集的隐私(例如恢复人脸识别系统中的人脸信息).如第2节所述,逆向攻击包括成员推理攻击(MIA)和属性推理攻击(PIA).在MIA中,攻击者可以推断训练数据集中是否包含特定的记录.在PIA中,攻击者可以推测训练数据集中是否存在一定的统计特征.最近的研究发现,在人口训练数据集中,某些阶层的人(如妇女和少数民族)的样本代表性不足,会影响最终模型的表现[23].模型逆向攻击表明,信息不仅可以从数据集流向模型和预测结果,还可以从模型和预测结果反向流向数据集.

现实中存在很多隐私风险,因此隐私保护是深度学习的关键.在训练过程中,用户不能自动删除公司收集的数据,不能控制自己如何使用数据,甚至不知道是否从数据中学习到了敏感信息.用户还承担着公司存储的数据被其他部门合法或非法访问的风险.在推理过程中,他们的预测数据和结果也会受到影响.模型提供者需要保护他们的模型和数据集不被公开.

在实施方面,隐私保护可以分为4种技术:1)差分隐私(DP-differential privacy)[6,24];2)同态加密(HE-homomorphic encryption)[25-26];3)安全多方计算(SMC-secure multi-party computation)[27-28];4)次优选择(SC-suboptimal choice)[8,29].

差分隐私是密码学中的一种手段,旨在最大限度地提高数据查询的准确性,同时尽可能减少从统计数据库[30]查询时识别其记录的机会.它主要通过删除个体特征并保留统计特征的方式来保护用户隐私.Dwork等人[31]首先提出了严格的数学定义,称为ε-indistinguishability和δ-approximate ε-indistinguishability,后来分别被称为ε-差分隐私和(ε,δ)-差分隐私.由于差分隐私在数据库中的应用,在深度学习中它经常被用来保护训练数据集的隐私.

同态加密是一种关注数据处理的加密技术,最早由Rivest在20世纪70年代提出,包括加法同态加密和乘法同态加密.Gentry在2009年首次设计了一个真正的全同态加密方案.同态加密是这样一种加密函数:对明文进行环上的加法和乘法运算,然后对其进行加密,和先对明文进行加密,再对密文进行相应的运算,可以得到等价的结果,即$En(x)\oplus En(y)=En(x+y)$.在深度学习中,同态加密通常被用来保护用户的预测数据和结果.一些工作也保留了训练模型的隐私.用户加密他们的数据并以加密的形式将其发送到MLaaS中,云服务将其应用于模型进行加密预测,然后以加密的形式返回给用户.

安全多方计算主要是为了在没有可信第三方的情况下,保证约定函数的安全计算,这始于百万富翁的问题.它主要采用的技术包括多方计算、加密电路和不经意传输.在深度学习过程中,其应用场景是多个数据方希望使用多个服务器对其联合数据进行模型训练.它们要求任何数据方或服务器不能从该过程中的任何其他数据方了解训练数据.安全多方计算可以保护训练数据集和训练模型.

与上述3种系统保护技术不同,次优选择是一种独特的保护方法.该方法易于实现,且具有较低的时间成本,但其效果尚未经过大规模实践的检验.例如,为防止盗窃模型参数,一些研究人员可能对模型参数进行四舍五入处理[29],将噪声添加到类概率[8],拒绝特征空间里的异常请求[10],返回第2或第3类的最大概率[7]等.所有这些方法都在一定程度上失去了一些准确性,从而换取隐私保护的改善.

因此,在应用这些防御技术之前,需要仔细考虑得失之间的平衡.

综上所述,我们将这些攻击目标与攻击方法和防御方法相关联.模型逆向攻击通常获取训练数据集的信息.模型提取攻击针对训练好的模型.预测数据和结果在传输过程中容易受到窃取等传统攻击.此外,差分隐私通常保护训练数据集,同态加密模型和预测数据,安全多方计算在训练过程中保护数据集和模型,次优选择主要针对训练模型.

3.3 攻击方法

本节详细介绍3种隐私攻击的技术方法.

3.3.1 模型提取攻击

模型提取攻击破坏了模型本身的隐私,攻击者试图窃取模型的参数和超参数.目前主流方法通过构建精确模型或相似模型来实现模型的提取.精确模型是指攻击者试图重建原始模型,或从原始模型计算参数或超参数;而相似模型是攻击者构建的一个在预测性能上相近的替代模型.窃取精确模型会损害模型所有者的核心商业资产,并为攻击者获取价值,而窃取相似模型通常用于生成可迁移的对抗样本.众所周知,对抗样本对深度学习是一个不小的威胁,但如果攻击者对模型一无所知,则很难生成可靠的对抗样本.通过发动模型提取攻击,攻击者以某种方式提取到原始模型、参数或结构,便可以利用这些知识来确定决策边界,从而生成相应的反例.

模型提取攻击的研究大多是在黑盒模型下进行的,在黑盒模型下只能得到训练模型的算法.攻击者通常构造特殊的输入,向预测API提交查询,并接收输出,获得许多输入输出对.由于训练后的机器学习模型本质上是一个函数,因此只要攻击者获得足够的输入输出对并有足够的时间,从理论上就可以恢复模型参数.实际上,攻击者需要做的是利用模型特性来生成包含更多信息的样本,以减少查询个数的需求和时间成本,有时甚至要牺牲一些准确性.

1) 精确模型.在模型的精确参数重构中,方程求解攻击方法在机器学习模型中具有良好的效果.Tramèr等人[32]介绍了一种通过预测API提取模型的方法.他们通过发送大量的查询建立了模型方程,并得到了相应的预测结果.但该方法仅适用于决策树、逻辑回归、简单神经网络等特定的机器学习模型,不适用于DNN.Wang等人[29]试图在已知模型算法和训练数据的前提下窃取超参数.超参数在文中称为λ,用于平衡目标函数中的损失函数和正则化项.由于训练过程要求目标函数最小,所以目标函数在模型参数处的梯度为0.根据这个性质,攻击者可以通过对模型的查询得到很多线性方程,即超参数、模型参数和输入数据之间的关系.最后,利用线性最小二乘法对超参数进行估计.Baluja等人[33]训练了一个名为元模型(meta model)的分类器来预测模型属性.攻击者将查询输入提交给目标模型,并将目标模型提供的输出作为元模型的输入,然后元模型尝试输出目标模型的属性.元模型可以推断系统架构、操作方法、训练数据集大小等信息.

2) 相似模型.相似模型只要求在模型的表现上与原模型近似,主要用于生成对抗样本等.Papernot等人[34]试图生成可迁移的、无目标的对抗样本.攻击者利用基于雅可比矩阵的数据集增强(Jacobian-based dataset augmentation, JbDA)技术生成合成样本来查询目标模型,并建立了一个近似于目标模型决策边界的攻击模型.然后攻击者利用攻击模型生成对抗样本,由于可移植性,这些样本会被目标模型误分类.Juuti等人[7]通过对DNN训练的正则化和对JbDA的一般化,提出了一种新的合成数据生成方法,生成了对抗样本.经过扩展后的JbDA技术在生成可迁移的有针对性的对抗样本和复制预测行为方面具有较高的效率.考虑到不同模型之间的差异,Papernot他们[35]还发现,关于目标模型体系结构的知识是不必要的,因为任何机器学习模型都可以用更复杂的模型来代替,比如DNN.

3.3.2 成员推理攻击

Truex等人[36]提出了MLaaS平台中成员推理攻击(membership inference attack, MIA)的一种通用的系统方案.给定实例x和对在数据集D上训练的分类模型F_t的黑盒访问权,当训练F_t时,对手是否能够在D中很有信心地推断实例x是否包含在D中.在MIA中,对手更关心x是否在D中,而不是x的内容.目前成员推理攻击可以通过3种方法实现:

1) 训练攻击模型.攻击模型是一个二元分类器,用来推断目标记录的信息.它将成员推理问题转化为分类问题,可用于白盒和黑盒攻击.很多研究还引入了影子模型来训练攻击模型,影子模型主要用来模拟目标模型,并生成攻击模型所需的数据集.当然,对影子模型的训练也会增加攻击代价.

Shokri等人[37]利用机器学习中的API调用,设计、实现并评估了黑盒模型的MIA攻击方法.他们生成了类似于目标训练数据集的数据集,并使用相同的MLaaS来训练影子模型.这些数据集是通过

基于模型的综合、基于统计的综合、有噪声的真实数据等方法得到的.使用影子模型为攻击模型提供训练集,训练集输入是某个记录的类标签、预测向量.输出是该记录是否属于影子模型训练集.训练好的攻击模型以类标签和预测向量作为输入,输出该记录是否在目标训练集中.Salem 等人[38]放宽了文献[37]中的部分约束条件(要在同一 MLaaS 上训练影子模型,影子模型和目标模型的数据集具有相同分布),并在没有目标模型的知识结构和训练数据集分布的情况下只使用一个影子模型.攻击模型以模型概率向量输出的前 3 个最大值作为输入确定隶属度.

Pyrgelis 等人[39]实现了在聚合位置数据上的 MIA.其主要思想是利用先验位置信息,通过具有识别功能的可识别博弈过程进行攻击.他们训练了一个分类器(即攻击模型)作为识别函数来确定数据是否在目标数据集中,无需影子模型.

2) 概率信息计算.该方法利用概率信息推断隶属度,无需攻击模型.举例来说,假设 a,b 都属于类别 A,其中 a 属于训练数据集而 b 不属于,由于 a 参与了训练过程,模型可能以 0.9 的置信概率将 a 分类为 A;考虑 b,由于它对模型而言是新出现的,尽管模型也能将 b 分类为 A,但可能只有 0.6 的置信概率.于是可以根据模型返回的最大类概率实施攻击.但这种方法需要一定的前提假设和辅助信息来获得可靠的概率向量或二元结果,这也是该方法在使用时的一个限制条件.

Fredrikson 等人[40]试图根据概率信息来构造某一数据是否出现在目标训练数据集中的概率.然后寻找概率最大的输入数据,得到的数据与目标训练数据集中的数据相似.Salem 等人[38]中的第 3 种攻击方法只需要记录通过目标模型输出的概率向量,并使用统计测量方法比较最大分类概率是否超过一个阈值,若超过则认为该记录属于数据集.Long 等人[41]提出了广义 MIA 方法,与文献[37]不同,它更容易攻击非过拟合数据.他们训练了大量类似于目标模型的参考模型(类似影子模型),根据参考模型的输出的概率信息选择易受攻击的数据,然后将目标模型和参考模型的输出进行比较,计算出数据属于目标训练数据集的概率.

3) 相似样本生成.该方法通过训练生成的模型(如生成对抗网络(GAN))生成训练记录,其生成的样本与目标训练数据集的样本相似.通过提高生成样本的相似度将使该方法更加有效.

Liu 等人[42]和 Hayes 等人[43]都探索了攻击生成模型的方法,不同于判别模型,生成模型通常用于学习数据的分布并生成相似的数据.文献[42]提出一种白盒攻击,用于单成员攻击和联合成员攻击.其基本思想是用目标模型训练生成的模型,以目标模型的输出为输入,以相似的目标模型输入为输出.训练后,攻击模型可以生成与目标训练数据集相似的数据.考虑到文献[37]中的方法难以攻击 CNN,Hitaj 等人[20]提出了一种更为通用的 MIA 方法,在协作深度学习模型的场景中执行了白盒攻击.他们构建了一个目标分类模型生成器,并利用该生成器形成了一个 GAN.经过训练后,GAN 可以生成与目标训练集相似的数据,但是这种方法的局限性在于,属于同一分类的所有样本都需要在视觉上相似,因此无法在同一个类别下区分它们.

3.3.3 属性推理攻击

属性推理攻击(property inference attack,PIA)是指对训练数据集的统计属性进行推理.推理的属性主要是一些统计信息,例如人口数据集中男女比例是否均衡、人口样本中是否存在少数民族样本、医疗数据集中患癌病人的比重等.

Ateniese 等人[44]首先提出了一种训练元分类器的白盒攻击方法.分类器以模型的特征信息作为输入,以训练该模型的数据集中是否包含特定属性为输出.他们还训练影子模型来为元分类器提供训练数据.由于他们主要提取机器学习模型的特征信息,这种方法在 DNN 上并不奏效.为了解决这个问题,Ganju 等人[45]构建了一个元分类器模型,该模型研究了如何提取 DNN 的特征值,使其作为元分类器的输入,其他部分与文献[44]非常相似.

另外,针对文献[20]中存在的不足,Melis 等人[46]提出了一种协作式学习的白盒攻击方法.其理论基础是,深度学习模型会记住太多数据特征[21].攻击者可以多次下载最新的模型,得到每个阶段的更新模型,减去不同阶段的聚合更新,并分析更新的信息来推断成员和属性.他们训练了一个二元分类器来判断数据集的属性,该分类器使用更新的梯度值作为输入.

3.4 防御方法

为了保护深度学习系统的隐私,一系列研究工作开发了不同的防御机制.基于对 27 篇文章的研究,我们将这些防御机制分为 4 类:差分隐私、同态加密、安全多方计算和次优选择.

3.4.1 差分隐私

差分隐私是一种密码学工具,旨在最大限度地提高数据查询的准确性,同时最大限度地减少查询统计数据库时识别其记录的机会.基于保护目标,差分隐私的方法可以从输出扰动、目标扰动和梯度扰动等方面进行扩展[24].这些方法分别指将随机扰动加到输出上、目标函数上和反向传播的梯度上.

Chaudhuri 等人[47]首先提出了输出和目标扰动,严格证明了凸损失函数机器学习模型中保持隐私,并将其实现为正则逻辑回归.输出扰动包括以边界灵敏度和增加灵敏度为基础的噪声训练模型.而 Wang 等人[24]表明,在非光滑条件下,输出扰动不能推广.Zhang 等人提出[48],在强凸的情况下,可以使用适当的学习率来提高操作速度和实用性.目标扰动是训练包含随机项的目标函数最小化的模型,它在理论上和经验上都优于输出扰动[49],但在实践中很难得到既保证隐私又保证效用的最优解.为了获得更好的性能或支持其他场景,Kifer 等人[50]通过选择高斯分布代替伽马分布提高了精度,并引入了第一个用于高维稀疏回归的差分隐私算法;文献[51]和文献[50]给出了 Lipschitz 损失函数的算法和证明.

Song 等人[52]提出了梯度扰动,其主要思想是在每次迭代更新参数时添加噪声.该方法不受强凸函数或强摄动优化问题的限制,在实际应用中具有一定的优越性.然而,由于随机梯度下降(SGD)或梯度下降(GD)的计算过程非常耗时,如果数据集很大,计算可能会花费很多时间.对于强凸前提,Bassily 等人[53]和 Talwar 等人[51]放宽了对 Lipschitz 凸函数的限制和严格的误差边界.然后,Abadi 等人[6]处理了非凸目标函数,并在适度隐私损失的情况下以适中的成本训练 DNN.他们对 DP-SGD 进行了修改和扩展,允许不同层的限幅阈值和噪声尺度不同.随后,Zhang 等人[48]首先给出了非凸优化问题的理论结果.文献[24]实现了满足 Polyak-Lojasiewicz 条件的非凸情况,并产生了更紧致的上界.Zhang 等人[54]与其他算法相结合,在分布式 ERM 中也应用了梯度扰动.

Hamm 等人[55]提出了一种使用局部分类器构造全局差分私有分类器的方法,该方法不需要访问任何一方的私有数据.Hynes 等人[56]提出了一种深度学习框架 Myelin,用于在可信硬件领域实现高效的、私有的、数据无关的实际深度学习模型.

目前已有一些度量标准被用来估计隐私风险.基于差分隐私的可组合性,最简单的度量方法是计算隐私消耗的总和[57].然而,直接将它们相加可能会得到松散的测量边界.文献[6]提出了一个更强的方法,主要采用标准 Markov 不等式来跟踪隐私损失,它在经验上获得了更严格的隐私损失约束.但上述指标仅限于 DP 框架,Long 等人[58]提出了差分训练隐私(DTP),可以测量不使用 DP 的分类器的隐私风险.

3.4.2 同态加密

一般的加密方案侧重于数据存储的安全性,而同态加密(HE)侧重于数据处理的安全性.HE 通常是在有数据泄漏风险中使用的.由于解密的高度复杂性,HE 可以有效地保护敏感数据不被解密和窃取.在深度学习中,它主要用于保护预测输入和结果,训练神经网络模型等.应用 HE 的主要负面影响是效率的降低,即错误传输问题、对密文的操作时间较长、加密后数据量急剧增加等.

Liu 等人[59]提出了 MiniONN,这是一个支持隐私保护的神经网络,并确保服务器不了解输入,客户端不了解模型.其主要思想是允许服务器和客户端为神经网络的每一层额外地共享输入和输出值.Jiang 等人[26]给出了一种矩阵和密码矩阵算术运算的实用算法.Phong 等人[4]提出了一个隐私保护 DL 系统,使用异步随机梯度下降应用于神经网络连接深度学习和密码学,并结合加性 HE.在其他方面,Hesamifard 等人[60]开发了 CryptoDL,用于在加密数据上运行 DNN,在 CIFAR-10 上的准确率为 91.5%.他们在 CNN 中利用低次多项式设计一个近似函数,然后用近似多项式代替原始的激活函数来训练 CNN,最后在加密数据上实现 CNN.

3.4.3 安全多方计算

在现实中,经常会遇到多个数据方希望在一台服务器上共同学习一个模型的场景.然而,每个数据方都不愿意将自己的数据共享给其他方.在多方数据只有一台服务器情况下,Shokri 等人[61]实现了一个系统,它允许多方在非共享输入数据集的情况下共同学习模型.各方都可以独立使用最终模型.在训练过程中,各个数据方对其局部数据集进行模型训练,再将所选参数的关键梯度上传到全局参数库,然后下载所需参数的最新值.基于这些性质,他们采用分布式选择性 SGD 方法来选择参数:1)梯度下降过程中不同参数的更新具有内在的独立性;2)不同的训练数据集对参数的贡献不同.Phong 等人[4]基于文献[61]进行了改进.每个数据方上传经过加法

HE 加密的梯度,并对模型应用异步 SGD.另外,Phong 等人[62]还提出了服务器辅助网络拓扑和全连接网络拓扑系统.各方共享神经网络的权值而不是梯度.他们不仅可以防范恶意的服务器,而且可以在即使只有一个诚实方的情况下防范数据方合谋.

在多方计算中另一个场景是,数据方不希望将所有训练数据交给一台服务器来训练模型.他希望将数据集分布到多个服务器,共同训练模型,每个服务器不会了解其他服务器的训练数据.SecureML[27]是一种保护隐私的双服务器模型协议.数据所有者将私有数据分配给 2 个非合谋服务器,并用安全的两方计算技术训练联合数据,支持安全的算术运算.采用了不经意传输和加密电路,并采用了面向多方计算友好的激活函数.Liu 等人[59]提出了一种支持隐私保护的神经网络 MiniONN.它确保服务器对输入一无所知,而客户机对模型一无所知.其主要思想是允许服务器和客户端额外地共享神经网络每一层的输入和输出值.

多方计算更一般化的场景是有 M 个数据方希望使用 N 台服务器对他们的联合数据进行模型训练.要求任何数据方或服务器对任何其他数据方的训练数据一无所知.在 SecureNN[28]里,$N=3$ 或 4,M 可以是任意值.此外,经过训练的模型作为一个秘密共享,并对任何单个服务器或数据端隐藏.这些秘密共享可以由服务器或任何其他方组合起来重构模型.

3.4.4 次优选择

为了抵御模型提取攻击,许多研究都试图在一定程度上向用户提供次优模型.Tramèr 等人[32]提出了第一个量化模型提取攻击预测概率的防御方法.他们只允许攻击者提取给定的类标签,而不提供置信度评分,或者提供四舍五入的置信度.该方法减少了向攻击者提供的信息量,但也减少了合法的服务.后来文献[7]表明,即使不使用预测概率,模型提取攻击也是有效的.但是 Lee 等人[8]发现,在类概率中注入噪声仍然可以延长攻击时间.攻击者被迫放弃概率信息,只使用标签信息,这大大增加了查询数量和攻击时间.Wang 等人[29]发现对模型参数进行四舍五入会增加攻击者对超参数攻击的估计误差.不幸的是,该误差对测试性能的影响可以忽略不计.

还有一种方法是从用户提交的查询请求中发现异常.Kesarwani 等人[10]依赖于记录客户端发出的所有请求,并计算正常请求组成的特征空间.当检测到新的请求空间超过预定阈值时,认为模型提取攻击发生.因此,他们需要在输入中对预测类进行线性分离来评估特征空间.此外,PRADA[7]是基于给定客户提交的样本分布的突然变化检测攻击,假设是攻击者提交的样本中的特征分布比良性查询中更不稳定.一旦 PRADA 检测到攻击,根据目标模型的预测,以最大概率返回第 2 类或第 3 类分类.PRADA 在检测对文献[34]的攻击时需要数百个查询,对文献[32]的攻击需要数千个查询.

3.4.5 其他方法

Xu 等人[63]将数据清洗以保护隐私.他们将原始数据用密码加密后发送给服务商.为了保护 MLaaS 中数据集的隐私,Zhang 等人[5]引入了一个混淆函数,并将其输入到模型训练任务中.混淆函数向现有样本添加随机噪声,或使用新样本增强数据集.因此,关于单个样本的特征或一组样本的统计特性的敏感信息是隐藏的.Nasr 等人[3]设计了一个 min-max 游戏,它在最小化模型的预测损失的同时,最大化推理攻击的收益,目标是共同最大化隐私和预测精度.

Cao 等人[64]提出了机器学习去除的思想,目标使机器学习模型完全忘记一段训练数据,并去除其对模型和特征的影响.它们将训练数据样本转化为一种求和形式,用来快速计算新模型.Hunt 等人[65]提出一个保护 SGX 上的 MLaaS 隐私的系统.它向服务运营商隐藏训练数据,既不向用户显示算法也不显示模型结构,只提供对训练模型的黑盒访问.Ohrimenko 等人[66]针对支持向量机、神经网络、决策树和 K-means 聚类问题,提出了一种基于 Intel Skylake 处理器的数据无关机器学习算法.

4 安 全

4.1 安全问题简介

安全与隐私在很多方面是密不可分的,但在这里,我们需要在人工智能领域区分安全与隐私问题.深度学习模型的形成依赖于对大量数据进行耗时耗力的训练.直观地说,训练数据、训练模型和预测输入都是所有者私有的,值得保护.众所周知,人工智能系统中已经存在着隐私研究对象,例如所收集的训练数据集、训练模型的参数、用户准备提交的预测数据以及模型返回的结果.要保护系统中原本存在的合法数据(模型参数、数据集等),就是隐私问题.

然而,在人工智能系统中,造成安全问题的恶意样本通常是未知的.例如投毒攻击将恶意数据添加

到训练数据集中,会对深度学习的预测产生负面影响.这些恶意样本不应该存在于其中.如何抵御这样的未知样本就是一个安全问题.此外,受到攻击的分类模型在训练过程中不会接触到这些对抗样本,这些恶意数据原本不在学习模型中.要防范系统中原本不存在的、可能引起模型出错的恶意数据,就是安全问题.

4.2 安全问题研究工作

在深度学习系统中,训练数据集和预测数据需要与用户交互,而训练过程和训练模型一般是封闭的.因此,训练数据集和预测数据更容易受到未知恶意样本的攻击.更具体地说,如果在训练数据集中出现恶意样本,我们称之为投毒攻击;如果在预测数据中出现恶意样本,我们称之为对抗攻击.我们共调查了89篇相关论文,其中15篇与投毒攻击相关,11篇与投毒防御相关,36篇与对抗攻击相关,27篇与对抗防御相关.

投毒攻击在训练过程中添加恶意样本,从而影响生成的模型.大多数恶意样本搜索方法都是通过发现算法或训练过程的漏洞来实现的.早期的机器学习算法也容易受到投毒攻击[67-69].投毒攻击主要在2个方面影响了正常模型.1)直接改变分类器的决策边界,破坏分类器的正常使用,使其不能正确地对正常样本进行分类,破坏了模型的可用性.这主要是通过错误标记数据实现的.攻击者使用错误的标签提交数据记录,或恶意修改训练数据集中现有数据的标签.2)在分类器中创建后门.它能正确地对正常样本进行分类,但会导致对特定数据的分类错误.攻击者可以通过后门进行有针对性的攻击,破坏模型的完整性.这主要是通过加入特定的数据实现的.它们向数据集提交包含特定特征(如水印)和标签的数据,而在其他数据记录中很可能没有这样的特征.此外,他们还可以直接攻击特征选择算法[70].相应地,防御方法主要是通过增强训练算法[71]的鲁棒性和保护数据集[72]的安全性来实现的.

在预测过程中,对抗攻击会对正常样本增加恶意干扰.对抗样本既要欺骗分类器,又要让人无法察觉.该攻击广泛应用于图像识别领域[73-76],也用于语音处理[77]、语音到文本转换[78]、文本识别[79]、恶意软件检测[80]等.目前,主流方法寻找扰动包括FGSM[73],JSMA[74],C&W[75],DeepFool[76],UAP[81],ATN[33]和一些变种.也有一些研究攻击了CNN,DNN之外的其他深度学习模型,甚至在现实世界中产生了对抗的实例.防御策略主要从对抗样本的生成和攻击的过程进行考虑,包括对抗训练[82]、基于区域的分类[83]、输入变化[84]、梯度正则化[85]、蒸馏[86]、数据处理[87]和训练防御网络[88].

4.3 投毒攻击

投毒攻击试图通过污染训练数据来降低深度学习系统的预测.由于它发生在训练阶段之前,通过调整相关参数或采用替代模型,所造成的污染是很难解决的.在机器学习的早期,投毒攻击被认为是对主流算法的一种重要威胁.例如,支持向量机[67,89-90]、贝叶斯分类器[68]、层次聚类[91]、逻辑回归[92]都受到了投毒攻击的危害.随着深度学习的广泛使用,攻击者也将他们的注意力转移到深度学习上了[93-95].

Muñoz-González等人[96]首先对基于反向梯度优化的多类问题进行了投毒攻击.该算法自动分步计算梯度,并对学习过程进行倒转,以降低攻击复杂度.通过添加一个投毒点,实现了通用或特定的错误攻击.这种攻击对许多深度学习任务都很有效,包括垃圾邮件过滤、恶意软件检测和手写数字识别.大多数投毒攻击研究集中在离线环境中,分类器在固定的输入上进行训练.然而,很多训练过程中数据以流的形式按顺序到达,即在线学习.Wang等人[97]对在线学习的数据投毒攻击进行了调查.他们将问题形式化为半在线和全在线2种设置,采用增量式、区间式和教学强化式3种攻击算法.他们的在线攻击比无视输入数据的在线特性的攻击要好.

综上所述,投毒攻击本质上是在训练数据上寻求全局或局部分布的扰动.众所周知,机器学习和深度学习的性能在很大程度上取决于训练数据的质量.高质量的数据通常应该是全面的、无偏见的和有代表性的.在数据投毒的过程中,错误的标签或有偏差的数据被有意地加工并添加到训练数据中,降低了整体质量.据观察,投毒有2方面原因:

1) 错误标记数据.在分类任务中,深度学习模型通常会在标记数据下提前进行训练.也就是说,$L:\{x_1,x_2,\cdots,x_n\}\rightarrow Y$,其中$Y$是给定输入的特定标签.通过将标签操作为$L:\{x_1,x_2,\cdots,x_n\}\rightarrow Y'$来生成错误标记的数据,其中$Y'$是一个错误的标签.错误标记数据的接受可能导致2种结果:深度学习不能有效地学习决策边界;将决策边界显著地推到不正确的区域.结果表明:该算法在容错条件下不能收敛.后者可以以相当小的损失终止,但是决策边界与正确边界之间的距离很大.

Xiao等人[90]通过翻转标签来调整训练集来攻击支持向量机,他们提出了一个优化的框架来寻找

标签翻转,使得分类误差最大化,从而降低了分类器的准确率.Biggio 等人[91]实现了针对单链接层次聚类的投毒攻击.它依靠启发式算法来寻找最优的攻击策略.他们使用模糊攻击来最大程度地降低聚类结果.Alfeld 等人[19]提出了一个在线性自回归模型下编码攻击者欲望和约束的框架.攻击者可以通过在训练数据中添加最优的特殊记录来将预测推向某个方向.Jagielski 等人[93]讨论了线性回归模型的投毒攻击.攻击者可以操纵数据集和算法来影响机器学习模型.他们引入了一种快速的统计攻击,这种攻击只需要有限的训练过程知识.

2) 特定混淆数据.机器学习实践者从大量信息中提取具有代表性的特征,用于学习和训练.这些特征的权重是经过训练确定的,对预测具有重要意义.然而,如果一些精心设计的数据具有无偏倚的特性分布,就会破坏训练,并得到一组误导性的特征权重.例如,将很多炸弹形状的图形标记为限速标志并将其放入数据集中学习,那么可能所有带有炸弹的图像将被标识为限速标志,即使它原本是一个停止(STOP)标志.

该方法在 LASSO, Ridge Regression, Elastic net 等特征选择算法中也很常见.Xiao 等人[70]直接研究了常见的特征选择算法在投毒攻击下的鲁棒性.结果表明,在恶意软件检测应用中,特征选择算法在投毒攻击下受到破坏性影响.通过插入少于 5% 的有毒训练样本,LASSO 特征选择过程得到的结果与随机选择几乎没有区别.Shafahi 等人[94]试图找到一个特定的测试实例来控制分类器的行为,而没有控制训练数据的标签.他们提出了一种水印策略,并训练了多个投毒的实例.在投毒实例中添加目标实例的低透明度水印,以允许某些不可分割的特性重叠.该方法为攻击者打开了一个分类器的后门,攻击者无需访问任何数据收集或标记过程.

4.4 对抗攻击

对抗攻击利用对抗样本(adversarial examples, AEs)使模型预测错误,也称为逃避攻击.对抗攻击是一种探索性攻击,它破坏了模型的可用性.AEs 是通过在原始样本中添加扰动而产生的.它们混淆了训练有素的模型,但在人类看来它们很正常,这保证了攻击的有效性.在图像处理中,通常使用小扰动来保证原样例与 AEs 之间的相似性.在语音和文本中,它确保 AEs 也是有意义的和上下文相关的.恶意软件检测保证 AEs 在添加扰动后仍具有原始恶意功能.

模型的误分类有目标性和非目标性两大类.前者要求 AEs 被错误地分类为特定的标签,以达到特殊的恶意目的.后者只要求 AEs 被错误分类(可以是任意错误标签),用于抵抗检测或其他场景.AEs 的生成过程通常需要最小化扰动,因为越小的扰动对人的影响也就越小.最小距离通常用 L_p 距离(或称 Minkowski 距离)来度量,常用的有 L_0,L_1,L_2 和 L_∞:

$$L_p(\boldsymbol{x},\boldsymbol{y})=\left(\sum_{i=1}^{n}|x^{(i)}-y^{(i)}|^p\right)^{\frac{1}{p}},$$

$$\boldsymbol{x}=(x^{(1)},x^{(2)},\cdots,x^{(n)}),$$

$$\boldsymbol{y}=(y^{(1)},y^{(2)},\cdots,y^{(n)}).$$

对抗攻击可以应用于许多领域,其中应用最广泛的是图像分类.通过添加小的扰动,我们可以生成对抗的图像,这些图像对人类而言很难区分,但是能造成模型的分类错误.对抗攻击也用在其他领域,比如音频[77,98]、文本[79]、恶意软件检测[99-101]等.Carlini 等人[78]提出了一种基于语音到文本神经网络的文本对抗攻击系统 DeepSearch.它可以通过添加小扰动将任意给定的波形转换成任意期望的目标短语.他们使用序列到序列的神经网络,产生超过 99.9% 的相似波形,并达到 100% 的攻击率.Gao 等人[79]提出框架 DeepWordBug 来在黑盒设置中生成对抗文本序列.他们使用不同的评分函数来处理更好的突变词.它们几乎最小化了编辑距离,并将文本分类精度从 90% 降低到 30%~60%.Rigaki 等人[102]使用 GANs 通过修改网络行为来模拟合法应用程序的流量来避免恶意软件检测.他们可以通过修改恶意软件的源代码来调整命令和控制(C2)通道来模拟 Facebook 聊天网络流量.最好的 GAN 模型在经过 300~400 个训练阶段后,每分钟产生一个以上的 C2 流量.文献[103-105]提出了在黑盒中生成恶意软件实例以进行攻击检测模型的方法.此外,文献[106]提出了一种针对二进制编码恶意软件检测的反攻击算法,实现了 91.9% 的准确率.

在图像领域,对抗攻击主要通过梯度下降法、最优化、神经网络自动化等方法搜索对抗样本来实现.一些研究也开始考虑现实世界中对抗样本的问题.在这里,我们定义 $F:\mathbb{R}^n\to\{1,2,\cdots,k\}$ 是将图像值向量映射到类标签的模型分类器.$Z(\cdot)$ 是倒数第 2 层的输出,通常表示类概率.$\boldsymbol{\delta}$ 是扰动,$\|\boldsymbol{\delta}\|_i$ 表示求 L_i 距离.我们接下来详细地介绍在图像领域里生成对抗扰动的 12 种方法.

1) L-BFGS 攻击.Szegedy 等人[1]提出盒约束的 L-BFGS,用于生成 AEs.他们还发现了 2 个与直觉

相反的特性.首先,该空间包含的语义信息位于神经网络的高层,而不是单个单元.其次,扰动或 AEs 具有较强的鲁棒性,可以在不同的神经网络或训练数据集之间共享.这些性质为今后的研究奠定了基础.

$$\min_{\boldsymbol{\delta}} c\|\boldsymbol{\delta}\|_2+Loss_F(\boldsymbol{x}+\boldsymbol{\delta},l),$$
$$\text{s.t.}\quad \boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n.$$

其中,l 是分类错误的标签,$\boldsymbol{x}+\boldsymbol{\delta}$ 是对抗样本,他们试图找到满足 $F(\boldsymbol{x}+\boldsymbol{\delta})=l$ 的 $\boldsymbol{\delta}$,要求扰动 $\boldsymbol{\delta}$ 尽量小,同时 $\boldsymbol{x}+\boldsymbol{\delta}$ 被分类为 l 的损失(即 $Loss_F(\boldsymbol{x}+\boldsymbol{\delta},l)$)也尽量小.损失函数满足 $Loss_F(\boldsymbol{x},F(\boldsymbol{x}))=0$,$c>0$ 是一个平衡 2 个最小值的超参数,$\boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n$ 保证添加扰动后的对抗样本仍在正常图像的取值范围内.

2) FGSM 攻击.FGSM(fast gradient sign method)是由 Goodfellow 等人[73]提出的.文章解释说,AEs 产生的原因是神经网络在高维空间中的线性行为,而不是非线性.设 l_x 是 $\boldsymbol{x}$ 的实际分类.损失函数描述输入 $\boldsymbol{x}$ 的损失.扰动 $\boldsymbol{\delta}$ 的方向是利用反向传播计算的梯度确定的.每个像素在梯度方向上的大小为 ε.随着 ε 的增加,扰动的大小和攻击成功率增加,被人发现的可能性也增加.

$$\boldsymbol{\delta}=\varepsilon\times\text{sign}(\nabla_x Loss_F(\boldsymbol{x},l_x)).$$

3) BIM 攻击.BIM(basic iteration method)[107]是 FGSM 的迭代版本,也称为 I-FGSM.$Clip_{x,\varepsilon}(\boldsymbol{x})$ 函数对每个像素的图像进行剪切,并使生成的 AE 在每次迭代时满足 L_∞ 的边界.I-FGSM 在白盒攻击中强于 FGSM,但其可移植性较差[108-109].

$$\boldsymbol{x}_0=\boldsymbol{x},$$
$$\boldsymbol{x}_{i+1}=Clip_{x,\varepsilon}(\boldsymbol{x}_i+\alpha\times\text{sign}(\nabla_x Loss_F(\boldsymbol{x}_i,l_x)).$$

4) MI-FGSM 攻击.MI-FGSM(momentum iterative FGSM)[110]是基于梯度引入的.Momentum 用于摆脱局部极值,迭代用于稳定优化.在白盒或黑盒模型上,该方法比基于梯度的单步法具有更强的可移植性.

$$\boldsymbol{x}_{i+1}=Clip_{x,\varepsilon}(\boldsymbol{x}_i+\alpha\times\frac{\boldsymbol{g}_{i+1}}{|\boldsymbol{g}_{i+1}|}),$$
$$\boldsymbol{g}_{i+1}=\mu\times\boldsymbol{g}_i+\frac{\nabla_x Loss_F(\boldsymbol{x}_i,y)}{\|\nabla_x Loss_F(\boldsymbol{x}_i,y)\|_1},$$

其中,y 是要被分类错误的目标类.与 BIM 不同的是,计算 $\boldsymbol{x}_{i+1}$ 时,不仅和当前损失函数的梯度方向有关,也和之前求出的损失函数(即 $\boldsymbol{g}_i$)有关.

5) JSMA 攻击.JSMA (Jacobian-based saliency map attack)[74]只改变了少量像素,而没有影响整个图像,它限制了 L_0 距离,而不是 L_2 和 L_∞.它们每次修改图像的个别像素,记录其对分类结果的影响,然后迭代地进行下去.对于任意一对像素 p,q,求解

$$\alpha_{pq}=\sum_{i\in\{p,q\}}\frac{\partial Z(\boldsymbol{x})_t}{\partial\boldsymbol{x}^{(i)}},$$
$$\beta_{pq}=\left(\sum_{i\in\{p,q\}}\sum_j\frac{\partial Z(\boldsymbol{x})_j}{\partial\boldsymbol{x}^{(i)}}\right)-\alpha_{pq}.$$

其中,α_{pq} 表示像素 p,q 对目标分类的影响,β_{pq} 表示对所有其他输出的影响.这张映射图上的值越大,意味着欺骗网络的可能性越大.

$$(p^*,q^*)=\arg\max_{(p,q)}(-\alpha_{pq}\times\beta_{pq})\times(\alpha_{pq}>0)\times(\beta_{pq}<0).$$

上式表明算法当前选择 (p^*,q^*) 像素对添加扰动从而实施攻击.(p^*,q^*) 满足:像素 p^*,q^* 对目标分类的影响为正,对所有其他输出的影响为负,且对目标分类的影响(α_{pq})以及对所有其他输出影响的绝对值($-\beta_{pq}$)两者乘积最大.

6) C&W 攻击.C&W[75]在 L_0,L_2 和 L_∞ 中实现了对蒸馏防御方法[111]的攻击.他们试图找到尽可能小的 δ,并欺骗分类器.与 L-BFGS 类似,C&W 主要优化了目标:

$$\min_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_p+c\times f(\boldsymbol{x}+\boldsymbol{\delta}),$$
$$\text{s.t. } \boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n.$$

$c>0$ 是一个超参数,用于控制 2 个目标函数之间的平衡.$f(\cdot)$是一个人工定义的函数,这里列举文中使用的函数:

$$f(\boldsymbol{x}+\boldsymbol{\delta})=\max(\max\{Z(\boldsymbol{x}+\boldsymbol{\delta})_i:i\neq t\}-Z(\boldsymbol{x}+\boldsymbol{\delta})_t,-K).$$

这里,$f(\cdot)\leqslant 0$ 当且仅当分类结果为对抗目标标签 t 时.K 保证 $\boldsymbol{x}+\boldsymbol{\delta}$ 将被高度信任地分类为 t.因此在最小化式子中,既要让扰动 $\boldsymbol{\delta}$ 尽量小,也要让 $f(\cdot)$尽量小,即分类结果为目标标签 t.C&W 保证生成的 AEs 一定会被错误分类,但由于计算量大,造成时间开销较大.

7) EAD 攻击.EAD (Elastic-net attacks to DNNs)[112]是用于制作 AEs 的弹性网络正则化攻击框架,它结合了 L_1,L_2 度量,提供了很少使用的面向 L_1 的样例,并将最好的 L_2 攻击作为一个特例.结果显示,EAD 设计的基于 L_1 的示例执行得和其他最佳攻击一样好.最优化公式为

$$\min_{\boldsymbol{\delta}} c\times f(\boldsymbol{x}+\boldsymbol{\delta})+\beta\|\boldsymbol{\delta}\|_1+\|\boldsymbol{\delta}\|_2^2,$$
$$\text{s.t. } \boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n,$$

其中,$f(\boldsymbol{x}+\boldsymbol{\delta})$与 C&W 中相同,$t$ 是目标标签.相较于 C&W,EAD 的优化公式中多了一个扰动项.显然,

当$\beta=0$时,C&W中的L_2目标攻击是EAD的一个特殊的情况.

8) OptMargin攻击.OptMargin[113]可以在有限的输入空间内躲避基于区域分类的防御.与以前的研究不同,它的目标是低维的子空间,不受空间周围邻居点的限制.该方法产生的AEs的判定边界与良性样本不同.然而,它无法模仿良性样本.OptMargin是C&W在L_2攻击的扩展,它在x的周围添加了许多目标函数.具体优化公式为

$$\min_{\boldsymbol{\delta}}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2+c\times(f_1(\boldsymbol{x})+\cdots+f_m(\boldsymbol{x})),$$

$$f_i(\boldsymbol{x})=\max(Z(\boldsymbol{x}+\boldsymbol{v}_i)_y-\max\{Z(\boldsymbol{x}+\boldsymbol{v}_i)_j:j\neq y\},-K).$$

这里,$\boldsymbol{x}_0$是原始样例.$\boldsymbol{x}=\boldsymbol{x}_0+\boldsymbol{\delta}$是对抗样本.$y$是$\boldsymbol{x}_0$的真实标签.$f_i(\boldsymbol{x})$是类似于C&W的目标函数,共$m$个.$\boldsymbol{v}_i$是应用于$\boldsymbol{x}$的扰动,共$m$个.OptMargin不仅保证对抗样本$\boldsymbol{x}$可以欺骗神经网络,还保证它周围的邻居$\boldsymbol{x}+\boldsymbol{v}_i$也可以.

9) DeepFool攻击.DeepFool[76]以迭代方式产生最小的归一化扰动.他们将图像逐步推入分类边界,直到符号发生变化.在相近的成功欺骗率下,DeepFool产生的扰动比FGSM要小.

10) NewtonFool攻击.NewtonFool[114]提出了一个强假设,即攻击者可以使用倒数第2层输出的类概率向量$Z(\boldsymbol{x})$.假设$l=F(\boldsymbol{x}_0)$,他们的目的是找到小的$\boldsymbol{\delta}$,使$Z(\boldsymbol{x}_0+\boldsymbol{\delta})_l=0$.他们用迭代方法把$Z(\boldsymbol{x}_0)_l$尽可能快地降到0.从$Z(\boldsymbol{x}_0)_l$开始,他们在每一步使用线性函数逼近新的$Z(\boldsymbol{x})_l$,即

$$Z(\boldsymbol{x}_{i+1})_l\approx Z(\boldsymbol{x}_i)_l+\nabla Z(\boldsymbol{x}_i)_l\times(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i),\ i=0,1,2,\cdots,$$

其中,$\boldsymbol{\delta}_i=\boldsymbol{x}_{i+1}-\boldsymbol{x}_i$是第$i$步迭代的扰动,最终扰动$\boldsymbol{\delta}=\boldsymbol{\delta}_0+\boldsymbol{\delta}_1+\cdots+\boldsymbol{\delta}_i$.结果表明它比FGSM,JSMA,DeepFool都快.

11) UAP攻击.UAP(universal adversarial pertur-bations)[81]可以以高概率在几乎任何输入数据上导致目标模型的错误分类.UAP对于数据和网络架构来说是通用的.让μ表示包含所有样例的数据集.它主要目的是寻找扰动$\boldsymbol{\delta}$,这个$\boldsymbol{\delta}$可以在几乎所有μ中的样本上欺骗$F(\cdot)$.

$$F(\boldsymbol{x}+\boldsymbol{\delta})\neq F(\boldsymbol{x}),\ \text{大部分}\ \boldsymbol{x}\in\mu,$$

其中扰动$\boldsymbol{\delta}$应满足约束条件:

$$\|\boldsymbol{\delta}\|_p\leqslant\xi,$$

$$P_{\boldsymbol{x}\in\mu}(F(\boldsymbol{x}+\boldsymbol{\delta})\neq F(\boldsymbol{x}))\geqslant 1-\lambda,$$

P表示概率,通常$0<\lambda\ll 1$.在没有优化或梯度计算的情况下,他们将每次迭代计算的最小扰动集合起来.Hayes等人[115]使用通用对抗网络(UANs)在有目标和无目标攻击中自动生成UAP.

12) ATN攻击.ATN[33]是一种训练有素的神经网络,可以高效、自动地攻击另一个目标.ATN通过添加最小扰动将任何输入转换为AE.他们使用有针对性的白盒ATNs来生成AEs,并成功地将83%~92%的图像输入转换为对ImageNet的对抗攻击.

13) 其他攻击方法.Papernot等人[34]提出了一种基于黑盒综合数据生成替代训练算法的新方法,在Google和Amazon上分别实现了96.19%和88.94%的准确率.Tramèr等人[116]提出了梯度对齐对子空间,它用于估计输入空间的未知维度.他们发现,子空间的很大一部分被2个不同的模型共享,从而实现了可移植性.他们首先寻找多个独立的攻击方向,定量研究模型决策边界的相似性.Narodytska等人[117]利用一种基于局部搜索的新技术构造了网络梯度的数值逼近,然后利用该技术构造了图像中的一组像素在黑盒中扰动.此外,Ilyas等人[118]引入了一个更加严格和实用的黑盒威胁模型.他们使用自然进化策略来执行黑盒攻击,减少了2~3个数量级的查询.

除了DNN外,还有很多研究人员对生成模型、强化学习和机器学习算法进行了深入的研究.Mei等人[119]为支持向量机、逻辑回归和线性回归确定最优训练集攻击.证明了最优攻击可以描述为一个双层优化问题,可以用梯度法求解.Huang等人[120]证明了对抗攻击策略在强化学习中也是有效的.Kos等人[121]对深度生成模型(如变分自编码器(VAE))进行了对抗攻击.他们的方法包括基于分类器的攻击,以及对潜在空间的攻击,这些攻击在MNIST,SVHN和CelebA上都表现得很好.

4.4.1 物理世界的对抗攻击

在图像识别领域,考虑到观察点、光照和相机噪声的影响,传统技术产生的AEs可能无法在物理世界中欺骗分类器.Kurakin等人[107]使用从手机摄像头拍摄的图像作为Inception v3图像分类神经网络的输入.结果表明,由原始网络构造的大量对抗图像,即使通过摄像机输入到分类器,也会产生误分类.Athalye等人[2]提出了一种EOT(expectation over transformation)算法,用于合成对物理世界具有鲁棒性的对抗样本.他们使用EOT的特殊应用,并通过3D渲染过程进行区分,从而生成对抗对象.结果表明,3D打印对象可以从各个角度欺骗现实世界的系统.

4.5 投毒防御

大多数的投毒攻击都集中在数据和算法上,因此防御方法主要考虑从保护数据和算法入手.

1) 保护数据.数据保护主要包括保护收集到的数据不受篡改、抵抗重写攻击、防止拒绝、防止数据伪造、检测有毒数据等[122-124].Olufowobi 等人[125]提出了一种物联网系统的数据来源模型,以提高数据的可信度和可靠性.该模型描述了创建或修改数据点的上下文.他们未来的工作是将该模型集成到物联网设备的数据源完整性检测算法中.Chakarov 等人[126]通过评价单个数据点对训练模型性能的影响,采用一种检测投毒数据的方法.他们需要通过比较可信数据集上的性能来评估模型.Baracaldo 等人[72]通过使用源信息作为过滤算法的一部分来检测投毒攻击.该方法提高了检测率.他们使用训练数据点的来源和转换上下文来识别有害数据.它是在部分可信和完全不可信的数据集上实现的.

2) 保护算法.学习算法总是要在防止正则化和减少损失函数之间做出权衡,这种不确定性可能导致学习算法的脆弱性.一些投毒攻击是根据自身的弱点来实施的,因此研究鲁棒机器学习算法是预防投毒攻击的有效途径.Candès 等人[127]首先研究了鲁棒 PCA 的鲁棒机器学习算法.它假定底层数据集的一小部分是随机销毁的,而不是有针对地销毁.Chen 等人[128]研究了对抗破坏下的鲁棒线性回归问题,Feng 等人[129]研究鲁棒的逻辑回归,他们都需要对特征独立性和亚高斯分布做出强有力的假设.Goodfellow 等人[73]提出了一种鲁棒线性回归方法,该方法放松了对特征独立性和低方差亚高斯噪声的假设,只假设特征矩阵可以用低秩矩阵逼近.该方法将鲁棒低秩矩阵近似与鲁棒主成分相结合,获得了较强的性能保证.Jagielski 等人[93]在训练过程中加入有毒的数据训练模型,而不是简单地删除它们.该方法迭代地估计回归参数,并将其训练在每次迭代中残差最小的点的子集上.本质上,它使用了一个根据每次迭代中残差的不同子集计算的被修剪的损失函数.

4.6 对抗防御

对抗攻击的防御方法主要从阻止对抗样本生成和检测对抗样本 2 个目标出发,本文总结了以下 7 种方法.

1) 对抗训练.对抗训练选择 AEs 作为训练数据集的一部分,使训练后的模型能够学习 AEs 的特征.Huang 等人[82]提出了一个较早的防御方法,即通过生成 AEs 作为中间步骤来学习具有强大对手的鲁棒分类器.同时他们也提出了一种新的 AEs 搜索方法.Kurakin 等人[109]将对抗训练应用于更大的数据集,如 ImageNet.其主要创新之处是批处理规范化、训练数据集(包括干净的和敌对的示例)和相对权重.他们还发现一步攻击比迭代攻击更具有可移植性.但是这种训练在正常样本上丧失了部分准确性.此外,集成对抗训练[108]包含了从其他预训练模型传输的每个输入.然而,对抗训练只能使训练模型对训练集中的 AEs 具有较强的鲁棒性,该模型不能学习训练集之外的 AEs 的特性.

2) 基于区域的分类.了解对抗样本区域的性质,并使用更健壮的基于区域的分类也可以抵御对抗攻击.Cao 等人[130]使用基于区域的分类(RC)代替基于点的分类开发了新的 DNNs.他们通过从以测试样本为中心的超立方体中随机选择几个点来预测标签.RC 将 C&W 攻击的成功率从 100%降低到 16%,但它对 OptMargin 攻击很难起作用.Pang 等人[83]使用了一种反向交叉熵防御方法.该分类器将正常样本映射到最终隐藏层空间的低维流形邻域.Ma 等人[131]提出了局部固有维数来表征对抗区域的维数特性.他们基于样本到邻域的距离分布,对样本区域的空间填充能力进行了评价.另外,Mccoyd 等人[132]在训练数据集中添加了大量不同类别的背景图像,以帮助检测 AEs.他们在 EMNIST 数据集中的关键类之间添加了背景类,背景类充斥在关键类之间的空白区域.该方法易于实施,但对 C&W 攻击没有效果.

3) 输入数据变换.改变或转换输入可以防御对抗攻击.Song 等人[84]发现 AEs 主要位于训练区域的低概率区域.因此他们设计了 PixelDefend,通过自适应地将 AE 向分布方向移动来净化 AE.Guo 等人[133]通过图像转换探索了图像分类系统的模型无关防御.他们的目的是消除输入的对抗扰动.他们的图像转换包括图像裁剪和重新缩放、位深度缩减、JPEG 压缩、总方差最小化和图像拼接.Xie 等人[134]在预测过程使用对输入的随机化来防御对抗攻击并减轻影响,包括随机调整图片大小和随机填充.该方法计算量小,与其他防御方法兼容.另外,Wang 等人[135]认为 AEs 比正常样本更敏感.如果将大量随机扰动添加到对抗样本和正常样本中,标签变化的比例会有显著差异,这样就可以识别对抗样本.他们在 MNIST 和 CIFAR-10 上实现了高准确度和低成本的差异判别.Tian 等人[136]认为 AEs 对某些图像

变换操作(如旋转和移位)比正常图像更敏感.他们用这种方法在图像分类中抵御了白盒的 C&W 攻击.Buckman 等人[137]提出了一种对神经网络进行简单修改的方法 TE(thermometer encoding).他们发现 TE 和热码离散化显著提高了网络对 AEs 的鲁棒性.

4) 梯度正则化.梯度正则化(或梯度掩蔽)是另一种有效的防御方法.Madry 等人[85]通过优化鞍点公式实现了这一点,鞍点公式包括由投影梯度下降(PGD)求解的内部最大值和由随机梯度下降(SGD)求解的外部最小值.但他们发现这不能保证在合理的时间内实现.Ross 等人[138]分析了输入梯度正则化,其目的是训练可微模型,以惩罚输入的微小变化.结果表明,输入梯度正则化增强了鲁棒性,与防御蒸馏和对抗训练有质的区别.

5) 防御蒸馏.Papernot 等人[86]提出了一种防御蒸馏方法.蒸馏主要是指将知识从复杂的结构转移到简单的结构中,从而降低 DNN 结构的计算复杂度.该方法能够成功地抑制 FGSM 和基于 Jacobian 的迭代攻击构造的 AEs.Papernot 等人[111]还利用防御蒸馏提取的知识对模型进行平滑处理,并降低了网络梯度的大小.网络梯度大意味着小的扰动会引起输出结果大的变化,有利于寻找对抗样本.

6) 数据处理.Liang 等人[139]引入标量量化和平滑空间滤波,以减小扰动的影响.他们使用图像熵作为度量标准,并对各种图像进行了自适应降噪.文献[87]中使用有界 ReLU 激活函数对冲对抗扰动的正向传播,并使用高斯数据增强方法增强泛化能力.Xu 等人[140]提出了基于特征压缩的反例检测方法,包括降低每个像素上颜色位的深度和空间平滑.

7) 防御网络.一些研究使用神经网络等工具对 AEs 进行自动对抗.Gu 等人[88]使用了带有收缩自编码器(CAEs)和去噪自编码器(DAEs)的深度收缩网络(DCN),它可以通过额外的噪声腐蚀和预处理去除大量的对抗噪声.Akhtar 等人[141]提出了一种微扰整流网络作为目标模型的预输入层,用于对抗 UAPs.它可以在不修改网络的情况下为已部署的网络提供防御,并抵御看不见的敌对干扰.MagNet[142]利用探测网络对远离流形边界的 AEs 进行探测,利用重整器对靠近边界的 AEs 进行改造.该过程不需要 AEs 或生成过程的知识.

5 总 结

随着人工智能领域在生活中各个方面的广泛应用[143-145],相关的安全问题也显现出来.本文调研了机器学习安全领域相关的 145 篇论文,并对机器学习系统所遇到的安全问题进行了完整而详细的划分.我们将该领域分为隐私和安全两大块,并按攻击目的、攻击目标、攻击过程将攻击分为 4 类.在每种攻击内部,按时间线和所采用的技术,将繁杂的研究进行总结归纳,划分了不同的技术,并对技术之间的优劣进行了比对和分析.在防御方面,我们着重保护机器学习系统的隐私和抵抗安全攻击,将每种防御类型内部的防御技术进行归类总结,并介绍了防御技术对攻击技术的适应性.另外,根据对这些攻击和防御技术的总结和研究,我们还提出了构建安全健壮的机器学习系统、保护机器学习所有参与者隐私安全的经验,也对目前机器学习系统以及人工智能领域的热点问题进行了讨论.

1) 提高数据质量,增强数据安全.机器学习在收集数据中可能会收集到脏数据,或者攻击者为实现投毒攻击所提供的数据,因此要对收集的数据做清洗,提高数据质量.一方面可以采用人工的方法对脏数据进行剔除,另一方面可以采用防御方法中对数据集进行清洗、保护的技术[5,63].面对数据量不足的情况,还可以通过构建生成模型(如 GAN)得到相似的数据.总之,训练的数据质量越高,训练获得的模型也越安全.

2) 保证个人数据隐私,防止模型滥用隐私信息.在目前的机器学习系统中,个人数据很难得到安全保障,模型可能从个人数据中推断出大量隐私信息.为保障用户的隐私安全,我们建议:引入监管部门对模型监控,严格监管模型对数据的使用,只允许模型提取允许范围内的特征,不可以擅自对敏感信息进行提取和推断;数据源保护,模型收集的数据必须进行去隐私化处理,模糊掉无关信息;建立健全相关法律法规,监管数据的收集、存储、使用和删除过程.

3) 通过模型解释性的研究解决模型安全性滞后性现状.目前来说,由于我们还没有实现对深度神经网络的深入理解(不清楚某条数据为什么预测出这个结果,不清楚不同数据对模型参数的影响程度),因此寻找安全问题进行攻击比提前预防要容易.因此我们亟需研究深度神经网络可解释性,尤其是 2018 年欧盟颁布了 GDPR 条例,更促进了神经网络可解释性的发展,相信随着对神经网络模型理解的加强,安全滞后性的问题将有效缓解.

4) 加强对人工智能在实际应用中的安全问题研究.人工智能的应用已经延伸至人类生活的物理

世界,如自动驾驶应用大量的图像识别技术.若其中存在安全问题将会直接造成对人身的物理伤害,从而导致了人们对 AI 安全的极大的恐慌.为了解决这个问题,我们要全面地研究机器学习系统易受的安全威胁,加强对模型的保护,加强对攻击方法的抵御.同时还要力求解释 AI 在何种情况下可能会出现状况、为什么会出现状况以及如何防止这种状况出现,并提出相应的防范措施,从而增强人们对 AI 技术应用的信任度.

总的来说,本文将机器学习系统所面临的安全问题进行了详细的分类,对未来的攻击防御技术的研究和发展有着重要意义.攻击和防御本身就是一场军备竞赛,对特定的攻击技术,可以研究专门的防御去抵抗它;而这种防御技术又会被其他的攻击技术所攻克.正是在这种攻防竞赛中,机器学习系统的安全性得以螺旋式的上升.在未来工作中,我们要继续研究机器学习领域的技术、应用和伦理方面的安全问题,并将模型提取攻击、模型逆向攻击、投毒攻击和对抗攻击中先进的攻防工作进行部署,从而对攻击和防御方法形成更统一和完整的度量.

参 考 文 献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [OL]. [2018-05-28]. http://arxiv.org/abs/1312.6199

[2] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples [C] //Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 284-293

[3] Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 634-646

[4] Phong L T, Aono Y, Hayashi T, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333-1345

[5] Zhang Tianwei, He Zecheng, Lee R B. Privacy-preserving machine learning through data obfuscation [OL]. [2018-05-28]. http://arxiv.org/abs/1807.01860

[6] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [C] //Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308-318

[7] Juuti M, Szyller S, Dmitrenko A, et al. PRADA: Protecting against DNN model stealing attacks [OL]. [2018-05-28]. http://arxiv.org/abs/1805.02628

[8] Lee T, Edwards B, Molloy I, et al. Defending against machine learning model stealing attacks using deceptive perturbations [OL]. [2018-05-28]. http://arxiv.org/abs/1806.00054

[9] Hua Weizhe, Zhang Zhiru, Suh G. Reverse engineering convolutional neural networks through side-channel information leaks [C] //Proc of the 55th IEEE Design Automation Conference. Piscataway, NJ: IEEE, 2018: 1-6

[10] Kesarwani M, Mukhoty B, Arya V, et al. Model extraction warning in MLaaS paradigm [OL]. [2018-05-28]. http://arxiv.org/abs/1711.07221

[11] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning [J]. Machine Learning, 2010, 81(2): 121-148

[12] Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety [OL]. [2018-05-28]. http://arxiv.org/abs/1606.06565

[13] Papernot N, McDaniel P, Sinha A, et al. Towards the science of security and privacy in machine learning [OL]. [2018-05-28]. http://arxiv.org/abs/1611.03814

[14] Bae H, Jang J, Jung D, et al. Security and privacy issues in deep learning [OL]. [2018-05-28]. http://arxiv.org/abs/1807.11655

[15] Papernot N, McDaniel P, Sinha A, et al. SoK: Security and privacy in machine learning [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2018: 399-414

[16] Liu Qiang, Li Pan, Zhao Wentao, et al. A survey on security threats and defensive techniques of machine learning: A data driven view [J]. IEEE Access, 2018, (6): 12103-12117

[17] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey [J]. IEEE Access, 2018, (6): 14410-14430

[18] Ling Xiang, Ji Shouling, Zou Jiaxu, et al. DeepSec: A uniform platform for security analysis of deep learning model [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 673-690

[19] Alfeld S, Zhu Xiaojin, Barford P. Data poisoning attacks against autoregressive models [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1452-1458

[20] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 603-618

[21] Song Congzheng, Ristenpart T, Shmatikov V. Machine learning models that remember too much [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 587-601

[22] Veale M, Binns R, Edwards L. Algorithms that remember: Model inversion attacks and data protection law [OL]. [2018-05-28]. http://arxiv.org/abs/1807.04644

[23] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification [C] // Proc of the Conf on Fairness, Accountability and Transparency. New York: ACM, 2018: 77-91

[24] Wang Di, Ye Minwei, Xu Jinhui. Differentially private empirical risk minimization revisited: Faster and more general [C] //Proc of the Annual Conf on Neural Information Processing Systems. New York: NIPS, 2017: 2719-2728

[25] Bachrach R, Dowlin M, Laine K, et al. CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy [C] //Proc of the 33nd Int Conf on Machine Learning. New York: ACM, 2016: 201-210

[26] Jiang Xiaoqian, Kim M, Lauter K, et al. Secure outsourced matrix computation and application to neural networks [C] // Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 1209-1222

[27] Mohassel P, Zhang Yupeng. SecureML: A system for scalable privacy-preserving machine learning [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 19-38

[28] Wagh S, Gupta D, Chandran N. SecureNN: Efficient and private neural network training [OL]. [2018-05-28]. https://eprint.iacr.org/2018/442

[29] Wang Binghui, Gong Zhenqiang. Stealing hyperparameters in machine learning [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2018: 36-52

[30] Proserpio D, Goldberg S, Mcsherry F. Calibrating data to sensitivity in private data analysis [J]. Proceedings of the VLDB Endowment, 2014, 7(8): 637-648

[31] Dwork C, Kenthapadi K, Mcsherry F, et al. Our data, ourselves: Privacy via distributed noise generation [C] //Proc of the 25th Annual Int Conf on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2006: 486-503

[32] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIs [C] //Proc of the 25th USENIX Security Symp. Berkeley, CA: USENIX Assocaiation, 2016: 601-618

[33] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 2687-2695

[34] Papernot N, Mcdaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning [C] //Proc of the 2017 ACM on Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519

[35] Papernot N, Mcdaniel P, Goodfellow I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples [OL]. [2018-05-28]. http://arxiv.org/abs/1605.07277

[36] Truex S, Liu Ling, Gursoy M E, et al. Towards demystifying membership inference attacks [OL]. [2018-05-28]. http://arxiv.org/abs/1807.09173

[37] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 3-18

[38] Salem A, Zhang Yang, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models [OL]. [2018-05-28]. http://arxiv.org/abs/1806.01246

[39] Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? membership inference on aggregate location data [C/OL] //Proc of the 25th Network and Distributed System Security Symp. Reston VA: The Internet Society, 2018 [2019-06-11]. https://arxiv.org/pdf/1708.06145.pdf

[40] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 2015 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[41] Long Yunhui, Bindschaedler V, Wang Lei, et al. Understanding membership inferences on well-generalized learning models [OL]. [2018-05-28]. http://arxiv.org/abs/1802.04889

[42] Liu Kin, Li Bo, Gao Jie. Generative model: Membership attack, generalization and diversity [OL]. [2018-05-28]. http://arxiv.org/abs/1805.09898

[43] Hayes J, Melis L, Danezis G, et al. LOGAN: Evaluating privacy leakage of generative models using generative adversarial networks [OL]. [2018-05-28]. http://arxiv.org/abs/1705.07663

[44] Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers [J]. International Journal of Security & Networks, 2015, 10(3): 137-150

[45] Ganju K, Wang Qi, Yang Wei, et al. Property inference attacks on fully connected neural networks using permutation invariant representations [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 619-633

[46] Melis L, Song Congzheng, Cristofaro E, et al. Inference attacks against collaborative learning [OL]. [2018-05-28]. http://arxiv.org/abs/1805.04049

[47] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression [C] //Proc of the 22nd Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2008: 289-296

[48] Zhang Jiaqi, Zheng Kai, Mou Wenlong, et al. Efficient private ERM for smooth objectives [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 3922-3928

[49] Chaudhuri K, Monteleoni C, Sarwate D. Differentially private empirical risk minimization [J]. Journal of Machine Learning Research, 2011, 12: 1069-1109

[50] Kifer D, Smith A, Thakurta A. Private convex optimization for empirical risk minimization with applications to high-dimensional regression [C] //Proc of the 25th Annual Conf on Learning Theory. Berlin: Springer, 2012: No.25

[51] Talwar K, Thakurta A, Zhang Li. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry [OL]. [2018-05-28]. http://arxiv.org/abs/1411.5417

[52] Song Shuang, Chaudhuri K, Sarwate A D. Stochastic gradient descent with differentially private updates [C] //Proc of the IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2013: 245-248

[53] Bassily R, Thakurta A. Private empirical risk minimization: ffficient algorithms and tight error bounds [C] //Proc of the IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2014: 464-473

[54] Zhang Tao, Zhu Quanqing. A dual perturbation approach for differential private ADMM-based distributed empirical risk minimization [C] //Proc of the 2016 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2016: 129-137

[55] Hamm J, Cao Yingjun, Belkin M. Learning privately from multiparty data [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 555-563

[56] Hynes N, Cheng R, Song D. Efficient deep learning on multi-source private data [OL]. [2018-05-28]. http://arxiv.org/abs/1807.06689

[57] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis [J]. Communications of the ACM, 2010, 53(9): 89-97

[58] Long Yunhui, Bindschaedler V, Gunter C. Towards measuring membership privacy [OL]. [2018-05-28]. http://arxiv.org/abs/1712.09136

[59] Liu Jian, Juuti M, Lu Yao, et al. Oblivious neural network predictions via miniONN transformations [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 619-631

[60] Hesamifard E, Takabi H, Ghasemi M. CryptoDL: Deep neural networks over encrypted data [OL]. [2018-05-28]. http://arxiv.org/abs/1711.05189

[61] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 2015 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310-1321

[62] Phong L T, Phuong T T. Privacy-preserving deep learning for any activation function [OL]. [2018-05-28]. http://arxiv.org/abs/1809.03272

[63] Xu Ke, Cao Tongyi, Shah S, et al. Cleaning the null space: A privacy mechanism for predictors [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 2789-2795

[64] Cao Yinzhi, Yang Junfeng. Towards making systems forget with machine unlearning [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 463-480

[65] Hunt T, Song Congzheng, Shokri R, et al. Chiron: Privacy-preserving machine learning as a service [OL]. [2018-05-28]. http://arxiv.org/abs/1803.05961

[66] Ohrimenko O, Schuster F, Fournet C, et al. Oblivious multi-party machine learning on trusted processors [C] //Proc of the 25th USENIX Security Symp. Berkeley, CA: USENIX Assocaiation, 2016: 619-636

[67] Bruckner M, Scheffer T. Nash equilibria of static prediction games [C] //Proc of the 22nd Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 171-179

[68] Nelson B, Barreno M, Chi F J, et al. Exploiting machine learning to subvert your spam filter [C] //Proc of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. Berkeley, CA: USENIX Assocaiation, 2008

[69] Mei Shike, Zhu Xiaojin. The security of latent dirichlet allocation [C] //Proc of the 8th Int Conf on Artificial Intelligence and Statistics. Cambridge, MA: MIT Press, 2015: 681-689

[70] Xiao Huang, Biggio B, Brown G, et al. Is feature selection secure against training data poisoning? [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2015: 1689-1698

[71] Liu Chang, Li Bo, Vorobeychik Y, et al. Robust linear regression against training data poisoning [C] //Proc of the 2017 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 91-102

[72] Baracaldo N, Chen B, Ludwig H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach [C] //Proc of the 2017 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 103-110

[73] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [OL]. [2018-05-28]. http://arxiv.org/abs/1412.6572

[74] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 372-387

[75] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57

[76] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE,2016: 2574-2582

[77] Yuan Xuejing, Chen Yuxuan, Zhao Yue, et al. CommanderSong: A systematic approach for practical adversarial voice recognition [C] //Proc of the 27th USENIX Security Symp. Berkeley, CA: USENIX Assocaiation, 2018: 49-64

[78] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 1-7

[79] Gao Ji, Lanchantin J, Soffa M L, et al. Black-box generation of adversarial text sequences to evade deep learning classifiers [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 50-56

[80] Kreuk F, Barak A, Aviv-Reuven S, et al. Deceiving end-to-end deep learning malware detectors using adversarial examples [C/OL] //Proc of the 32nd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018 [2019-06-11]. https://arxiv.org/abs/1802.04528

[81] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 86-94

[82] Huang Ruitong, Xu Bing, Schuurmans D, et al. Learning with a strong adversary [OL]. [2018-05-28]. http://arxiv.org/abs/1511.03034

[83] Pang Tianyu, Du Chao, Zhu Jun. Robust deep learning via reverse cross-entropy training and thresholding test [OL]. [2018-05-28]. http://arxiv.org/abs/1706.00633

[84] Song Yang, Kim T, Nowozin S, et al. PixelDefend: Leveraging generative models to understand and defend against adversarial examples [OL]. [2018-05-28]. http://arxiv.org/abs/1710.10766

[85] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [OL]. [2018-05-28]. http://arxiv.org/abs/1706.06083

[86] Papernot N, Mcdaniel P. On the effectiveness of defensive distillation [OL]. [2018-05-28]. http://arxiv.org/abs/1607.05113

[87] Zantedeschi V, Nicolae M I, Rawat A. Efficient defenses against adversarial attacks [C] //Proc of the 2017 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 39-49

[88] Gu Shixiang, Rigazio L. Towards deep neural network architectures robust to adversarial examples [OL]. [2018-05-28]. http://arxiv.org/abs/1412.5068

[89] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines [C] //Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2012

[90] Xiao Han, Xiao Huang, Eckert C. Adversarial label flips attack on support vector machines [C] //Proc of the 20th European Conf on Artificial Intelligence. Ohmsha, Japan: IOS, 2012: 870-875

[91] Biggio B, Pillai I, Bulo S R, et al. Is data clustering in adversarial settings secure? [C] //Proc of the 2013 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2013: 87-98

[92] Mei Shike, Zhu Xiaojin. Using machine teaching to identify optimal training-set attacks on machine learners [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 2871-2877

[93] Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 19-35

[94] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks [OL]. [2018-05-28]. http://arxiv.org/abs/1804.00792

[95] Steinhardt J, Koh P W, Liang Pang. Certified defenses for data poisoning attacks [C] //Proc of Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 3520-3532

[96] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization [C] //Proc of the 2017 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 27-38

[97] Wang Yizhen, Chaudhuri K. Data poisoning attacks against online learning [OL]. [2018-05-28]. http://arxiv.org/abs/1808.08994

[98] Gong Yuan, Poellabauer C. Crafting adversarial examples for speech paralinguistics applications [OL]. [2018-05-28]. http://arxiv.org/abs/1711.03280

[99] Huang Wenyi, Stokes J W. MtNet: A multi-task neural network for dynamic malware classification [C] //Proc of Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2016: 399-418

[100] Papernot N, Mcdaniel P, Swami A, et al. Crafting adversarial input sequences for recurrent neural networks [C] //Proc of the IEEE Military Communications Conf. Piscataway, NJ: IEEE, 2016: 49-54

[101] Pascanu R, Stokes J W, Sanossian H, et al. Malware classification with recurrent networks [C] //Proc of the IEEE Int Conf on Acoustics, Speech and Signal. Piscataway, NJ: IEEE, 2015: 1916-1920

[102] Rigaki M, Garcia S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 70-75

[103] Hu Weiwei, Tan Ying. Black-box attacks against RNN based malware detection algorithms [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 245-251

[104] Hu Weiwei, Tan Ying. Generating adversarial malware examples for black-box attacks based on GAN [OL]. [2018-05-28]. http://arxiv.org/abs/1702.05983

[105] 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 [C] //Proc of the 21st Int Symp on Research in Attacks, Intrusions and Defenses. Berlin: Springer, 2018: 490-510

[106] Al-Dujaili A, Huang A, Hemberg E, et al. Adversarial deep learning for robust detection of binary encoded malware [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 76-82

[107] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world [OL]. [2018-05-28]. http://arxiv.org/abs/1607.02533

[108] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [OL]. [2018-05-28]. http://arxiv.org/abs/1705.07204

[109] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale [OL]. [2018-05-28]. http://arxiv.org/abs/1611.01236

[110] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9185-9193

[111] Papernot N, Mcdaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 582-597

[112] Chen P Y, Sharma Y, Zhang Huan, et al. EAD: Elastic-net attacks to deep neural networks via adversarial examples [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 10-17

[113] He W, Li Bo, Song D. Decision boundary analysis of adversarial examples [C/OL] //Proc of Int Conf on Learning Representations. 2018 [2019-06-11]. https://openreview.net/pdf?id=BkpiPMbA-

[114] Jang U, Wu Xi, Jha S. Objective metrics and gradient descent algorithms for adversarial examples in machine learning [C] //Proc of the 33rd Annual Computer Security Applications Conf. New York: ACM, 2017: 262-277

[115] Hayes J, Danezis G. Learning universal adversarial perturbations with generative models [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 43-49

[116] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples [OL]. [2018-05-28]. http://arxiv.org/abs/1704.03453

[117] Narodytska N, Kasiviswanathan S. Simple black-box adversarial attacks on deep neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1310-1318

[118] Ilyas A, Engstrom L, Athalye A, et al. Query-efficient black-box adversarial examples (superceded) [OL]. [2018-05-28]. http://arxiv.org/abs/1712.07113

[119] Mei Shike, Zhu Xiaojin. Using machine teaching to identify optimal training-set attacks on machine learners [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 2871-2877

[120] Huang S, Papernot N, Goodfellow I, et al. Adversarial attacks on neural network policies [OL]. [2018-05-28]. http://arxiv.org/abs/1702.02284

[121] Kos J, Fischer I, Song D. Adversarial examples for generative models [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 36-42

[122] Wang Xinlei, Zeng Kai, Govindan K, et al. Chaining for securing data provenance in distributed information networks [C] //Proc of the 31st IEEE Military Communications Conf. Piscataway, NJ: IEEE, 2012: 1-6

[123] Lyle J, Martin A P. Trusted computing and provenance: Better together [C/OL] //Proc of the 2nd Workshop on the Theory and Practice of Provenance. Berkeley, CA: USENIX Association, 2010 [2019-06-11]. https://www.usenix.org/legacy/event/tapp10/tech/full_papers/lyle.pdf

[124] Hasan R, Sion R, Winslett M. The case of the fake picasso: Preventing history forgery with secure provenance [C] //Proc of the Conf on File and Storage Technologies. Berkeley, CA: USENIX Assocaiation, 2009: 1-14

[125] Olufowobi H, Engel R, Baracaldo N, et al. Data provenance model for internet of things (IoT) systems [C] //Proc of Int Conf on Service-oriented Computing. Berlin: Springer, 2016: 85-91

[126] Chakarov A, Nori A, Rajamani S, et al. Debugging machine learning tasks [OL]. [2018-05-28]. http://arxiv.org/abs/1603.07292

[127] Candès E J, Li Xiaodong, Ma Yi, et al. Robust principal component analysis? [J]. Journal of the ACM, 2011, 58(3): 11.1-11.37

[128] Chen Yudong, Caramanis C, Mannor S. Robust high dimensional sparse regression and matching pursuit [OL]. [2018-05-28]. http://arxiv.org/abs/1301.2725

[129] Feng Jiashi, Xu Huan, Mannor S, et al. Robust logistic regression and classification [C] //Proc of Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 253-261

[130] Cao Xiaoyu, Gong N Z. Mitigating evasion attacks to deep neural networks via region-based classification [C] //Proc of the 33rd Annual Computer Security Applications Conf. New York: ACM, 2017:278-287

[131] Ma Xingjun, Li Bo, Wang Yisen, et al. Characterizing adversarial subspaces using local intrinsic dimensionality [OL]. [2018-05-28]. http://arxiv.org/abs/1801.02613

[132] Mccoyd M, Wagner D. Background class defense against adversarial examples [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 96-102

[133] Guo Chuan, Rana M, Cisse M, et al. Countering adversarial images using input transformations [OL]. [2018-05-28]. http://arxiv.org/abs/1711.00117

[134] Xie Cihang, Wang Jianyu, Zhang Zhishuai, et al. Mitigating adversarial effects through randomization [OL]. [2018-05-28]. http://arxiv.org/abs/1711.01991

[135] Wang Jingyi, Sun Jun, Zhang Peixin, et al. Detecting adversarial samples for deep neural networks through mutation testing [OL]. [2018-05-28]. http://arxiv.org/abs/1805.05010

[136] Tian Shixin, Yang Guolei, Cai Ying. Detecting adversarial examples through image transformation [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 4139-4146

[137] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples [C/OL] //Proc of the Int Conf on Learning Representations. 2018 [2019-06-11]. https://openreview.net/pdf?id=S18Su--CW

[138] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1660-1669

[139] Liang Bin, Li Hongcheng, Su Miaoqiang, et al. Detecting adversarial image examples in deep networks with adaptive noise reduction [C/OL] //Proc of the IEEE Transactions on Dependable and Secure Computing. Piscataway, NJ: IEEE, 2018: 1-1 [2019-06-11]. https://arxiv.org/pdf/1705.08378.pdf

[140] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [C/OL] //Proc of the 25th Network and Distributed System Security Symp. Reston, VA: The Internet Society, 2018 [2019-06-11]. https://arxiv.org/pdf/1704.01155.pdf

[141] Akhtar N, Liu Jian, Mian A. Defense against universal adversarial perturbations [OL]. [2018-05-28]. http://arxiv.org/abs/1711.05929

[142] Meng Dongyu, Chen Hao. MagNet: A two-pronged defense against adversarial examples [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 135-147

[143] Zhao Qianqian, Chen Kai, Li Tongxin, et al. Detecting telecommunication fraud by understanding the contents of a call [J/OL]. Cybersecurity, 2018 [2019-06-11]. https://cybersecurity. springeropen. com/articles/10. 1186/s42400-018-0008-5

[144] Chen Yi, Zha Mingming, Zhang Nan, et al. Demystifying hidden privacy settings in mobile apps [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 570-586

[145] You Wei, Zong Peiyuan, Chen Kai, et al. SemFuzz: Semantics-based automatic generation of proof-of-concept exploits [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 2139-2154

He Yingzhe, born in 1995. PhD. His main research interests include machine learning, artificial intelligence security, and adversarial attack.

Hu Xingbo, born in 1996. Master. Her main research interests include software engineering and machine learning. (huxinbo@iie.ac.cn)

He Jinwen, born in 1997. PhD. Her main research interests include AI security and code analysis. (hejinwen@iie.ac.cn)

Meng Guozhu, born in 1987. PhD from the Nanyang Technological University, Singapore in 2017. Research Fellow at Nanyang Technological University. Visiting Research Fellow at the University of Luxembourg. Associate professor with the Institute of Information Engineering, Chinese Academy of Sciences. His main research interests include mobile security, big data analysis, vulnerability detection, program analysis, and machine learning. (mengguozhu@iie.ac.cn)

Chen Kai, born in 1982. PhD from the University of Chinese Academy of Science in 2010. Professor with the Institute of Information Engineering, Chinese Academy of Sciences. Professor with the University of Chinese Academy of Sciences. His main research interests include software analysis and testing; smartphones and privacy. (chenkai@iie.ac.cn)