计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200920
58(5):909-926,2021

针对深度学习模型的对抗性攻击与防御

李明慧[1,2] 江沛佩[1,2] 王 骞[1,2] 沈 超[3,4] 李 琦[5]
[1](空天信息安全与可信计算教育部重点实验室(武汉大学) 武汉 430072)
[2](武汉大学国家网络安全学院 武汉 430072)
[3](智能网络与网络安全教育部重点实验室(西安交通大学) 西安 710049)
[4](西安交通大学电子与信息学部 西安 710049)
[5](清华大学网络科学与网络空间研究院 北京 100084)
(minghuili@whu.edu.cn)

Adversarial Attacks and Defenses for Deep Learning Models

Li Minghui[1,2], Jiang Peipei[1,2], Wang Qian[1,2], Shen Chao[3,4], and Li Qi[5]
[1](*Key Laboratory of Aerospace Information Security and Trusted Computing*, *Ministry of Education* (*Wuhan University*), *Wuhan* 430072)
[2](*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)
[3](*Key Laboratory for Intelligent Networks and Network Security* (*Xi'an Jiaotong University*), *Ministry of Education*, *Xi'an* 710049)
[4](*Faculty of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)
[5](*Institute for Network Sciences and Cyberspace*, *Tsinghua University*, *Beijing* 100084)

Abstract Deep learning is one of the main representatives of artificial intelligence technology, which is quietly enhancing our daily lives. However, the deployment of deep learning models has also brought potential security risks. Studying the basic theories and key technologies of attacks and defenses for deep learning models is of great significance for a deep understanding of the inherent vulnerability of the models, comprehensive protection of intelligent systems, and widespread deployment of artificial intelligence applications. This paper discusses the development and future challenges of the adversarial attacks and defenses for deep learning models from the perspective of confrontation. In this paper, we first introduce the potential threats faced by deep learning at different stages. Afterwards, we systematically summarize the progress of existing attack and defense technologies in artificial intelligence systems from the perspectives of the essential mechanism of adversarial attacks, the methods of adversarial attack generation, defensive strategies against the attacks, and the framework of the attacks and defenses. We also discuss the limitations of related research and propose an attack framework and a defense framework for guidance in building better adversarial attacks and defenses. Finally, we discuss several potential future research directions and challenges for adversarial attacks and defenses against deep learning model.

收稿日期:2020-11-09;**修回日期**:2021-02-22
基金项目:国家重点研发计划项目(2020AAA0107700);国家自然科学基金优秀青年科学基金项目(61822207);国家自然科学基金重点项目(U20B2049)
This work was supported by the National Key Research and Development Program of China (2020AAA0107700), the National Natural Science Foundation of China for Excellent Young Scientists (61822207), and the Key Program of the National Natural Science Foundation of China (U20B2049).
通信作者:王骞(qianwang@whu.edu.cn)

Key words artificial intelligence security; deep learning; adversarial attack; defense strategy; privacy protection

摘 要 以深度学习为主要代表的人工智能技术正在悄然改变人们的生产生活方式,但深度学习模型的部署也带来了一定的安全隐患.研究针对深度学习模型的攻防分析基础理论与关键技术,对深刻理解模型内在脆弱性、全面保障智能系统安全性、广泛部署人工智能应用具有重要意义.拟从对抗的角度出发,探讨针对深度学习模型的攻击与防御技术进展和未来挑战.首先介绍了深度学习生命周期不同阶段所面临的安全威胁.然后从对抗性攻击生成机理分析、对抗性攻击生成、对抗攻击的防御策略设计、对抗性攻击与防御框架构建4个方面对现有工作进行系统的总结和归纳.还讨论了现有研究的局限性并提出了针对深度学习模型攻防的基本框架.最后讨论了针对深度学习模型的对抗性攻击与防御未来的研究方向和面临的技术挑战.

关键词 人工智能安全;深度学习;对抗性攻击;防御策略;隐私保护

中图法分类号 TP391

随着移动互联网飞速发展、硬件设备持续升级、海量数据产生和算法不断更新,人工智能(artificial intelligence, AI)的发展已势不可挡,正逐渐渗透并深刻地改变着人类生产生活.深度学习(deep learning, DL)技术及其应用的发展令人瞩目,使强人工智能离人类生活越来越近.目前,基于深度学习的智能技术被广泛应用在人机交互、视觉处理、智能决策、自治系统、推荐系统、安全诊断与防护等各个领域.

目前,以深度学习为主要代表的人工智能开始进入产业化开发与深耕阶段,促进了各个领域的深刻变革.例如,深度学习驱动的数据分析技术已经从根本上改变了现有的视频监控、医疗健康和金融管理等系统的开发和应用.在安全领域,最新的检测防护系统能够利用深度学习技术从大规模数据资源中快速准确地提取出有用的可执行信息.

尽管深度学习被认为是深刻改变人类社会生活、改变世界的颠覆性技术,但是与任何一种先进技术发展和应用的过程类似,当面向用户的服务越来越成熟,客户资源逐渐增长,最终安全性会成为进一步广泛部署人工智能系统的最大挑战.

以深度学习为代表的人工智能技术,至今仍然是一个黑匣子.目前对深度学习模型的内在脆弱性以及针对其弱点设计的对抗性攻击技术的理解尚不充分,需要基础理论揭示深度学习背后的机理.但是深度学习模型参数规模大、结构复杂、可解释性差,对于对抗性攻击的生成机理分析十分困难.针对这些问题,目前已有相关研究尝试对各种对抗性攻击的作用机理进行解释[1-41].其次,深度学习框架在软件实现中不断暴露出新的漏洞,对抗性攻击的恶意样本生成和训练数据的污染致使系统形成漏判或者误判[42-77],甚至导致系统崩溃或被劫持.与此对应,目前研究者们有针对性地提出了多种防御方法[78-119],如模型隐私保护、模型鲁棒性增强以及输入样本对抗噪声检测与擦除等,以增强深度学习模型的抗攻击能力.同时,从系统的角度考虑,深度学习模型的攻击、防御的研究不能是单一、碎片化的,需要建立深度学习模型的对抗性攻击和防御框架,目前这方面的研究还处于起步阶段[120-146].

由于不同学者所处的研究领域不同,解决问题的角度不同,针对深度学习模型攻击与防御研究的侧重点不同,因此亟需对现有的针对人工智能的隐私保护的研究工作进行系统地整理和科学地分析、归纳和总结.在本文中,我们首先介绍了深度学习模型生命周期与安全威胁,然后从对抗性攻击生成机理、对抗性攻击生成方法、对抗攻击的防御策略、对抗性攻击与防御框架4个角度对现有的深度学习模型攻击与防御方法进行系统地总结和科学的归纳,并讨论了相关研究的局限性.最后,我们在现有基础上提出了针对深度学习模型攻击与防御的基本框架,并展望了深度学习模型对抗性攻防的未来研究方向.

1 深度学习生命周期与安全威胁

深度学习模型是当前人工智能系统大爆发的核心驱动,主要包括训练和推理2个阶段.在训练阶段,首先构建训练数据集,然后利用训练集对模型参数

进行训练调节,得到深度学习模型.当训练完成时,深度学习模型就进入了推理阶段,首先获取输入样本,然后将样本输入模型进行推理,得到相应的模型预测判别结果.

图 1 展示了深度学习的整个生命周期中模型潜在的安全威胁,包括训练集构建、模型训练、输入样本获取和模型应用 4 个攻击面的脆弱性和典型攻击方法.

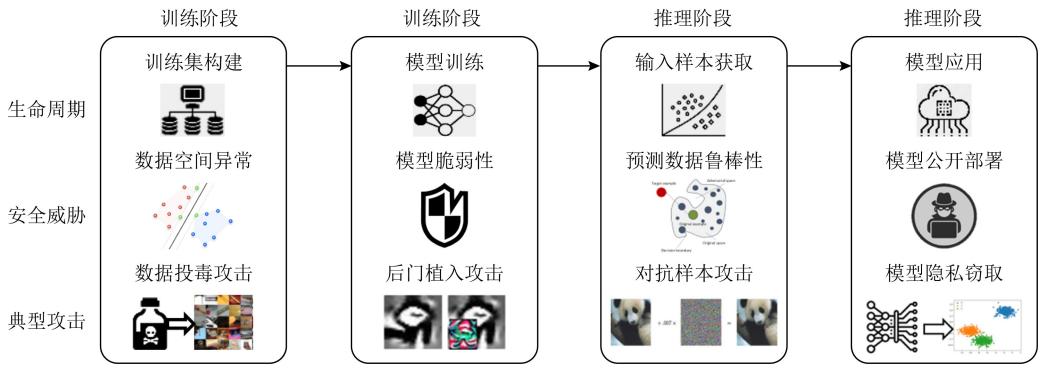


Fig. 1 The life cycle and security threats of deep learning models

图 1 深度学习模型生命周期与安全威胁

1) 训练集构建.在数据集构建阶段,需要收集大量数据用于训练,训练集的质量对深度学习模型至关重要.在这一过程中,攻击者可以通过注入精心构造的污染数据来导致训练数据空间异常,从而破坏、操控模型的可用性,使模型异常分类.典型的攻击手段有数据投毒攻击.

2) 模型训练.在模型训练阶段,需要根据训练数据对深度学习模型参数进行训练和调节.由于深度学习网络具有脆弱性,攻击者可以通过数据污染和修改模型参数等方法误导模型的训练过程,改变训练模型导致模型针对特定样本分类出错.典型的攻击手段有后门植入攻击.

3) 输入样本获取.在输入样本获取阶段,需要得到推理和预测的输入样本数据.然而,深度学习模型具有一定的脆弱性,攻击者在输入样本中添加精心构造的微小扰动,便可误导模型错误分类.典型的攻击手段有对抗样本攻击.

4) 模型应用.在模型应用阶段,用户可以根据需要访问模型.这一过程存在着模型信息泄露的威胁,如果攻击者知道其相关信息就可以有针对性地进行攻击.典型的攻击手段有模型窃取攻击、模型逆向攻击等.

本文将从对抗性攻击生成机理、对抗性攻击生成方法、对抗攻击的防御策略以及对抗攻击与防御框架 4 个角度分别针对训练和推理 2 个阶段的 4 个攻击面的相关工作进行阐述、总结,并分析现有工作的不足.

2 对抗性攻击生成机理分析

为更好地理解对抗性攻击的存在机理,推动通用性防御技术的发展,已有相关研究尝试对各种对抗性攻击的作用机理进行解释.表 1 对比和总结了现有对抗性攻击生成机理研究的代表性工作.

Table 1 The Summary of Representative Generation Mechanisms of Adversarial Attacks Methods

表 1 对抗性攻击生成机理研究典型方法对比与总结

攻击阶段	分析层面	机理概要	相关工作
训练阶段	数据层面	数据空间异常	文献[1-3,36]
	模型层面	模型脆弱性	文献[4-7]
推理阶段	数据层面	训练数据有限	文献[10-12,15]
		数据分布不一致	文献[13-14,37]
		数据高维特性	文献[16-19]
	模型层面	模型线性解释	文献[20]
		决策空间异常	文献[24-25]
		提取特征能力不足	文献[22-23,26-28,33-35]
		学习算法不当	文献[21,29-32]

2.1 训练阶段对抗性攻击生成机理

针对训练阶段的对抗性攻击生成机理研究,主要是对投毒攻击(poisoning attack)和后门攻击(backdoor attack)的作用机理进行研究.投毒攻击和后门攻击的主要思想是通过将恶意数据注入到训练数据中,使模型无法有效学习决策边界或学习到错误的决策边界.众所周知,深度学习的性能在很大

程度上取决于训练数据质量,而高质量数据通常应是全面、无偏见和有代表性的.

因此,基于数据层面分析,相关研究[1-3]表明投毒攻击的本质原理是通过在训练数据上寻求全局或局部分布扰动,使训练数据的整体分布发生畸变从而影响模型训练.

另外,基于模型层面分析,部分研究结果将其归咎于学习算法和模型脆弱性,导致学习阶段无法提取到具有代表性的特征,最终使模型过于关注具有误导性的特征而发生训练错误[4-7].

后门攻击通过对恶意数据加入特定标志的触发器,使神经网络训练时建立起部分神经元与触发器之间的联系,以致模型识别到触发器后,部分神经元激活异常,最终导致输出错误结果[8-9].

目前,针对训练阶段的对抗性攻击生成机理研究较少且主要集中于传统机器学习模型,其相关理论无法适用于更加复杂的深度学习模型.因此,针对深度学习训练阶段的对抗性攻击生成机理亟需深入探索和研究.

2.2 推理阶段对抗性攻击生成机理

针对推理阶段的对抗性攻击生成机理研究,主要以对抗样本为主.对抗样本通过对原始样本加入精心构造的微小噪声便能使深度学习模型识别出错.

现有部分研究尝试从数据空间角度进行机理解释.例如,McDaniel 等人[10]的研究表明:对抗样本产生的原因主要是训练模型的样本集不可能覆盖所有可能性,所以不可能训练出一个覆盖所有样本特征的模型,这就导致训练后模型的决策边界与真实决策边界不一致,而这两者之间的差异就是对抗样本空间.文献[11-12]也显示神经网络在训练时只学习到了一些局部子区域,而对抗样本往往处于流形空间的低概率区域,超出了模型学习概率分布的所在支集.同时,文献[13-14]指出对抗样本分布与真实样本分布不一致,所以对抗样本的误分类只是一种正常的测试误差现象.因此,Schmidt 等人[15]指出,构建鲁棒的模型比"标准"模型需要更多的训练数据.另外,文献[16-19]指出输入空间的高维特性也是对抗样本产生的重要因素.

与此同时,部分研究工作尝试从模型角度进行理论分析.Goodfellow 等人[20]提出,模型在高维空间的线性特性是对抗样本存在的真正原因.但 Tanay 等人[21]指出该线性解释存在诸多局限性,并表明当分类边界靠近数据流形时会存在对抗样本.文献[22-23]认为对抗样本的存在是由于分类器的"低灵活性",导致对抗样本不仅影响深度神经网络,而且影响所有分类器.Tabacof 等人[24]将对抗样本产生的原因归咎于分类器分类时所构建的稀疏和不连续的流形空间.He 等人[25]也指出对抗样本往往处于模型决策空间的畸形区域.另外,文献[26-28]研究表明神经网络过于关注不具有代表性的特征,因而很容易受到对抗信息的干扰.此外,部分研究指出学习算法的不足也是导致神经网络学习异常的原因,例如不恰当的交叉熵损失函数[29-30]、不充分的正则化等[21,31-32].同时,为更好地理解深度学习模型将对抗样本分类错误的现象,一些研究基于模型可解释性并可视化模型中间特征来直观地感受模型提取特征的差异,尝试解释其对抗机理[33-35].然而,相关研究工作并未给出成熟、统一的理论结果.

对抗性攻击生成机理的研究同时也推动着防御技术的发展,包括从输入空间分布及维度进行检测防御[2,36-37]、从模型角度来提升模型鲁棒性等[5-6,38-39].然而,这些防御方法具有很强的局限性,往往无法抵御更新的攻击方法[40-41].现有大多数对抗性攻击生成机理分析方法具有一定的片面性.

3 对抗性攻击生成方法

随着深度学习的快速发展,其内在的安全隐患也逐渐被暴露出来.针对深度学习模型的对抗性攻击主要分为训练阶段的攻击和推理阶段的攻击.

3.1 训练阶段对抗性攻击生成方法

在模型训练阶段,攻击者可以通过数据投毒攻击以及以此为基础发展的后门攻击技术来操控模型的可用性,使模型分类异常.针对模型训练,现有代表性对抗性攻击方案的对比和总结如表 2 所示:

Table 2 The Summary of Representative Adversarial Attacks Methods in the Training Stage

表 2 训练阶段对抗性攻击典型方法对比与总结

攻击类型	模型参数	训练数据	应用领域	相关工作
数据投毒	需要	需要	文本	文献[43]
	需要	需要	图像	文献[44]
	不需要	不需要	图像	文献[3]
后门攻击	需要	需要	图像	文献[45]
	需要	需要	图像,语音	文献[9]
	需要	不需要	图像,语音	文献[8,46]
	不需要	不需要	图像	文献[47]

在模型训练中,数据投毒攻击可通过篡改训练数据的内容和分布等方式来污染训练样本,最终影响模型训练效果.根据不同的攻击方式,现有方案主要分为2类:模型偏斜(model skewing)和反馈武器化(feedback weaponization).模型偏斜的主要攻击目标是训练数据样本,通过污染训练数据达到改变模型分类器边界的目的.反馈武器化的主要攻击目标是人工智能学习模型本身,利用模型自我训练和用户反馈机制发起攻击,使模型性能与预期产生偏差.自1993年数据中毒思想[42]被提出之后,Nelson等人[43]首次实现了数据投毒攻击方法,即针对基于贝叶斯学习的二分类算法成功攻击垃圾邮件分类与异常检测系统.然而该攻击方法局限性较大,无法适用于其他分类算法.Biggio等人[44]将污染模型训练转化为利用Karush-Kuhn-Tucker方法的数学优化问题,实现了对支持向量机(support vector machine, SVM)的攻击,方法泛化性更强.最近,Muñoz-González等人[3]首次在多分类问题中使用投毒攻击,提出基于反向梯度优化的攻击方法.结果显示,针对包含深度学习模型等在内的一系列基于梯度方法训练的模型,都可实现数据投毒效果.然而,当前研究的主要问题在于其没有针对深度神经网络的投毒攻击进行广泛评估,而深度网络的鲁棒性导致投毒攻击的效果较差.

进一步地,攻击者可以通过投毒攻击来添加后门实现后门攻击.其核心思想是将精心构造的恶意数据样本(带有错误的标签,通常是特定类别的标签)放到训练集中,破坏原有训练数据的概率分布,使训练出的模型在特定后门触发器被触发时产生分类错误,从而完成攻击.Gu等人[45]对神经网络模型的参数和训练数据进行修改,从而在神经网络模型上安插后门,导致在具有特定后门触发器的输入出现时,神经网络模型会错误分类,而正常输入情况下模型分类正常.Liu等人[8]在不能接触到训练集的情况下,通过模型逆向技术先生成替代训练集,然后完成后门攻击.Ji等人先后在文献[9,46]中分析了机器学习系统各部分组件的工作原理,实现了仅能操控部分系统结构假设条件下的后门攻击.在文献[47]中,Chen等人假设存在一个内部的攻击者,其通过污染训练集向模型注入后门.然而,现有针对神经网络的后门攻击威胁模型假设过强,在实际场景下很难实现.

3.2 推理阶段对抗性攻击生成方法

在模型推理阶段,攻击类型主要有2种:1)在输入样本获取阶段针对输入样本的攻击,即对抗样本攻击(adversarial examples attack);2)在模型应用阶段针对模型隐私的攻击,包括模型提取攻击(model extraction attack)和模型逆向攻击(model reverse attack).表3对比和总结了现有模型隐私窃取攻击的一些典型方法.

Table 3 The Summary of Representative Model Privacy Stealing Attacks Methods

表3 模型隐私窃取攻击典型方法对比与总结

攻击类型	攻击场景	查询数量	应用领域	相关工作
模型提取	黑盒	适中	自然	文献[48-49]
	黑盒	适中	合成	文献[50]
	黑盒	高	合成	文献[51]
模型逆向	白盒	高	自然	文献[52,55-56]
	黑盒	高	自然	文献[57]

模型提取攻击考虑基于隐私数据训练得到的模型,目的是提取该模型的各种属性信息或者构造与之功能相同的替代模型.在早期研究中[48],攻击者仅能通过大量的公开数据集去查询目标模型来获得每个样本相应的输出,继而构造替代模型.然而,此类方法仅适用于简单的机器学习模型如逻辑回归(logistic regression).针对神经网络模型,Orekondy等人[49]提出利用强化学习技术,有目的地选取某些标签下的数据集作为查询集,构造功能相似的神经网络替代模型.Juuti等人[50]在对训练数据分布有一定了解的情况下,改进超参数选择和人工样本生成过程从而增强攻击能力.在文献[51]中,Duddu等人在实施攻击之前,首先使用侧信道攻击来推测深度神经网络的深度,从而提高替代模型与目标模型的相似度.然而,使用当前的模型提取方法构造一个准确率较高的替代模型仍需大量的查询,易被防御检测系统发现.

进一步地,模型提取攻击可以作为推断模型训练集信息即模型逆向的基础.模型逆向攻击主要分为成员推理攻击(membership inference attack)和模型反演攻击(model inversion attack).成员推理攻击是一种针对人工智能系统的直接攻击.2008年,Homer等人[52]开发了一种针对基于统计学的机器学习模型的攻击技术.该攻击方法在文献[53-54]中得到进一步研究,主要用来比较关于此数据集(特别是次要等位基因频率)发布的统计数据和这些统计数据在一般人群中的分布情况,从而推断某一数据集中是否存在特定基因.Calandrino等人[55]提出可

根据特定的基于协同过滤推荐系统的时间行为进行攻击.攻击者利用协同推荐系统的输出变化来推断引起这些变化的输入.可以看到,使用成员推理攻击可以获得训练集的一些特征,但为更好地辅助对抗样本攻击,这些特征并不充分,仍然需要攻击者获得更多的训练集信息.模型反演是利用模型的输出去推断输入(一般是某个隐藏输入)的某些特征.Fredrikson 等人[56]详细分析了模型反演,并解释了模型逆向的基本工作原理,然后在文献[57]中首次提出利用预测产生的置信度实现模型反演,并将模型反演应用于面部识别模型.与成员推理不同,模型反演不会产生任何特定训练集中的图像.如果一个类中图像是多样的(例如类中可能包含多个个体或许多不同对象),那么模型反演的结果在语义上将会是无意义的,且不能被识别为来自训练集的任何特定图像.总之,模型反演只是产生了可以最好地描绘整个输出类特征的平均值.它既不能构造训练集中的特定成员,也不能在给定一个输入和一个模型的情况下,确定该特定输入是否曾用于训练模型.然而,由于模型逆向攻击获取了模型训练集的不同特征和信息,可为设计对抗样本攻击提供支持,有助于实现进一步的攻击.

深度神经网络推理阶段的另一个重要攻击手段是对抗样本攻击.一个典型例子即 Szegedy 等人[11]在 2013 年所描述的视觉"对抗样本"现象:针对输入图片构造肉眼难以发现的轻微扰动,可导致基于深度神经网络的图像识别器输出错误结果.根据攻击者可获得关于该模型的信息量,对抗样本攻击可分为白盒攻击(white-box attack)[58]和黑盒攻击(black-box attack)[59].表 4 总结了当前主流对抗样本攻击方法.

白盒攻击指攻击者完全知道分类模型和训练方法,同时可以访问训练数据分布以及完全训练的模型架构参数.相反地,黑盒攻击无法获取模型信息,只能通过一些设置或过去的输入信息来分析模型脆弱性.现有对抗样本攻击方法大都通过一个优化过程来生成对抗样本[58].2015 年,Goodfellow 等人[20]首次提出利用快速梯度符号法(fast gradient sign method, FGSM),将构建对抗性图像的非凸优化问题近似转化为线性形式,但该算法无法保证攻击成功率,尤其在定向攻击中成功率较低.随后,Kurakin 等人[60]提出通过多轮迭代来改进快速梯度符号法,但代价是增加了计算量.Tramèr 等人[61]提出随机快速梯度符号法(randomized fast gradient sign method),主要思想是在梯度计算之前对输入样本添加随机扰动,跳出数据点附近的大曲率,使快速梯度符号法生成的对抗样本更具泛化能力.2016 年,Moosavi-Dezfooli 等人[62]提出另外一种基于决策超平面的对抗样本生成技术,利用目标模型的迭代线性表示来生成对抗样本.由于该方法是限定神经网络下的超平面问题,因此在非神经网络模型上的应用相当有限,算法的通用性较低且需要大量时间分析模型特性.然而,本节讨论的攻击方法以及文献[41,63-64],都是属于白盒攻击.在更为普遍的场景中,当被攻击者使用了某些防御机制时,这些攻击方法就会失效.此外,攻击者通常难以全面了解模型信息,因此白盒攻击方法的通用性与实用性不高.

Table 4 The Summary of Representative Methods of Adversarial Example Generation

表 4 对抗样本生成典型方法对比与总结

攻击场景	攻击方法	攻击意向	攻击效率	样本失真	可迁移性	相关工作
白盒	损失梯度	非定向	高	是	否	文献[20]
		定向	低	是	是	文献[61]
	基本迭代	定向	高	是	是	文献[60]
	基于优化	定向	低	否	是	文献[41]
	决策面	非定向	低	否	否	文献[62-63]
黑盒	可转移性	非定向	高	是	是	文献[66-67]
	零阶优化	定向	低	是	否	文献[68]
		定向	高	否	否	文献[69]
	演化算法	非定向	低	是	否	文献[70]
		非定向	高	是	是	文献[73]
	决策边界	非定向	低	否	否	文献[71-72]
	生成式对抗网络	定向	低	是	否	文献[74-75]
		定向	低	否	否	文献[76-77]

更为广泛、实际的对抗样本攻击通常基于黑盒场景[65].黑盒攻击可分 3 种类型:基于可转移性、基于预测分值和基于决策标签的攻击.基于可转移性(transferability)的攻击是指攻击者试图构造一个替代模型(substitute model)来模拟被攻击模型,然后利用白盒攻击方法生成对抗样本攻击替代模型,最终基于对抗样本的可转移性攻击黑盒模型[66-67].然而,攻击替代模型通常会导致更大失真和较低成功率.于是,研究者们考虑基于预测分值的攻击,即通过模型得到样本预测概率.现有主流方法依赖于近似梯度来生成对抗样本,比如 Chen 等人[68]提出不需要梯度信息的零阶优化(zeroth order optimization

Zoo)算法实现有效黑盒攻击;在此基础上,Tu 等人[69]通过引入基于自动编码器的方法和自适应随机梯度估计(adaptive random gradient estimation)来平衡查询计数和失真优化的查询复杂性;在文献[70]中,预测分值是通过硬标签来估计的,然后利用自然进化策略(natural evolutionary strategies, NES)使目标类的概率最大化或样本原始类别的概率最小化.更严格的黑盒攻击是基于决策标签的攻击(decision-based attack),即只能得到预测的硬标签;2018 年 Brendel 等人[71]提出一种边界攻击(boundary attack)方法,其思想是基于在决策边界上的随机游动来生成对抗样本;类似地,Thomas 等人[72]解决了只知决策标签黑盒设置下查找通用扰动的问题;陈等人[73]提出一种基于精英策略的非支配遗传算法(non-dominated sorting genetic algorithm, NSGA),用黑色扰动块代替淤泥实现环境鲁棒的车牌误识别,但同时也带来一定的样本失真.总的来说黑盒攻击方法存在的共性问题是通常需要通过大量的查询来生成具有最小扰动的对抗样本或者收敛到具有少量查询的大扰动.还有一些基于生成对抗网络(generative adversarial networks, GAN)的对抗样本生成方法[74-77],然而这类方法的攻击效率较低.Hu 等人[74]利用原始样本和黑盒模型产生的输出来训练一个辨别器,再用辨别器指导生成器产生对抗样本.然而,该算法强调整个样本集的正确率,而不关心单个样本的相似度.这意味着每个样本所需的扰动可能很大,生成对抗样本的相似度低,在实际攻击场景中有较大局限性.

4 对抗攻击的防御策略

为抵御多样的对抗性攻击,其针对性的防御方案同样引发了研究者们的高度关注和重点研究.表 5 对比和总结了现有一些典型的对抗攻击防御方案.

Table 5 The Summary of Representative Defenses Methods Against Adversarial Attacks

表 5 现有对抗攻击防御的典型方法对比与总结

防御的攻击类型	部署阶段	防御方法	模型依赖	攻击方法依赖	相关工作
数据投毒	训练集构建	数据检测	否	否	文献[78-80,83]
后门植入			否	否	文献[81-82]
隐私窃取	模型训练	隐私保护	否	否	文献[93-101]
传感器欺骗	输入样本获取	传感器信号防护	否	否	文献[107-109]
对抗样本欺骗	模型训练	模型增强	否	否	文献[20,86,89]
			否	是	文献[87-88]
	模型推理应用	对抗样本检测	否	否	文献[110-111,117]
			否	是	文献[115]
			是	否	文献[116]
		对抗噪声擦除	否	否	文献[112-113,119]
			否	是	文献[118]
			是	否	文献[114]

4.1 线下阶段对抗攻击的防御策略

为从线下训练数据集中过滤掉恶意样本,保证训练集不被污染,Baracaldo 等人[78-79]最早提出将数据溯源技术用于深度学习系统,依靠反复训练模型来确定训练集中的异常数据.然而当训练集数据量较大时,该防御方案的检测效率较低.而异常检测技术只需分析数据本身的特征即可确定异常数据.Liu 等人[80]利用支持向量机和决策树(decision tree)来检测异常数据,指出这种技术需要保证在训练检测器时使用的是真实样本.为了克服检测器被污染的问题,Steinhardt 等人[81]通过构建各种后门攻击的近似上界,使检测器能在被污染的情况下检测异常数据.然而,当数据集中被污染的数据占比过大时,该防御方案会失效.Chen 等人[82]用聚类算法来判别异常数据,巧妙地回避了检测器污染问题.但该工作只讨论了数据污染,未考虑标签污染.Paudice 等人[83]提出一种算法来检测并重新标记被污染的标签,但仅考虑了二分类问题.目前,将数据溯源技术应用于深度学习系统可靠数据集构建的研究还不充分,且单独依靠数据溯源技术很难提供可靠数据集.

而基于数据异常检测的数据集清洗方法,缺乏有效反馈机制,需要反复清洗被污染的数据源数据.

可靠的数据集可以帮助提升模型的表现能力,消除某些潜在安全隐患,但对模型鲁棒性提升有限,还需要依靠优化模型构造过程来提升模型防御能力.Goodfellow 等人[20]提出通过对抗训练的方法来优化模型参数,增强模型鲁棒性.但 Kurakin 等人[84-85]指出对抗训练能够增强模型对单步攻击的鲁棒性,但很难抵抗迭代攻击,而且同样无法抵御利用单步攻击方法从另一个脆弱性模型中生成的扰动.另外,一些研究者通过优化网络结构来提升模型鲁棒性.Gu 等人[86]提出在网络输入层之前添加一个去噪自编码器以降低对抗噪声,但可能导致网络更容易遭受攻击[41].Lee 等人[87]使用一种常见的生成对抗网络架构来训练一个可以防御快速梯度符号法攻击的模型,其缺点是模型训练过程比较复杂.Bradshaw 等人[88]提出一种高斯混合深度模型,利用高斯过程对不确定因素的处理能力来增强模型的鲁棒性,并证明其对快速梯度符号法攻击的抵御能力.还有一些研究者通过简化已训练完成的模型结构来消除潜在威胁.Papernot 等人[89]提出利用神经网络知识蒸馏(knowledge distillation)技术来抵御对抗性攻击,但这种方法的防御能力有限[41].目前,已知现有线下训练模型鲁棒性增强策略只能防御已有攻击手段,且对已有攻击也无法进行全方位防御,另外,线下模型增强策略灵活性较差,不能根据已有模型的脆弱性进行动态调整.纪等人[90]指出,利用模型可解释性来分析和调试模型的错误决策行为,诊断模型中存在的缺陷,可为模型缺陷修复提供支撑,以获得更加鲁棒的模型.然而,现有模型诊断策略[91-92]缺乏与实际模型训练过程的有机结合.

对一个鲁棒的模型来说,如果攻击者知道其相关信息就可以有针对性地进行攻击.因此,保护模型的隐私信息同样被认为是提升模型防御能力的重要环节.为达到此目的,Graepal 等人[93]利用同态加密(homomorphic encryption)技术在加密数据上进行训练,但在加密数据上进行乘法操作会引入大量噪声,导致信息无法解密.因此,该方案需对现有算法进行最高次有界的多项式逼近,而这又会导致性能下降.文献[94-95]利用差分隐私(differential privacy)的思想,通过在训练模型时往参数中添加噪声的方式,达到隐私保护目的.但这些方法未考虑添加噪声对模型可用性的影响.文献[96-97]将训练过程当作是可用性和隐私保护之间的平衡优化问题来进行模型训练,但如果同时需要考虑模型鲁棒性,模型训练过程可能会很复杂.针对已训练好的模型,文献[98-99]利用同态加密的方法对模型加密,但同态加密引入的噪声会显著降低模型表现性能.Tramèr 等人[100]和 Wang 等人[101]分别提出对模型预测结果和模型参数做近似处理,以抵御模型窃取攻击,这些方法的防御效果还需要进一步进行验证.现有深度学习隐私保护方案主要通过线下训练掩盖真实模型参数的方式来实现隐私保护,多以牺牲模型可用性(例如运行时间、预测准确率等)为代价.

4.2 线上阶段对抗攻击的防御策略

另有一些研究重点关注模型发布之后的线上防御阶段.研究者们提出通过利用传感器增强方案在数据输入阶段过滤掉对抗性噪声,但目前大部分工作都停留在理论分析阶段[102-106],不过也有研究者提出了切实可行的传感器防护措施[107-109].在软件层面,有些研究利用对抗性噪声的不稳定性在预测任务时检测或擦除对抗性噪声.文献[110-111]通过对数据压缩或是添加噪声的方法试图破坏对抗样本的对抗性,然后比较处理前后数据预测结果是否发生改变来检测对抗样本.文献[112-113]通过 JPEG 压缩,去除图像方块中的高频信号成分,从而消除对抗性噪声.Xie 等人[114]在分类网络之前增加 2 个随机变化层,破坏对抗性噪声的特定结构来实现防御.然而,这些方法未对数据处理对真实样本预测结果的影响进行深入研究.为消除数据处理操作对真实样本的影响,Metzen 等人[115]训练了一个二分类检测网络,以模型某一隐藏层的输出作为输入,再输出此次输入样本为对抗样本的概率,但该方案只能检测特定类型的攻击.Feinman 等人[116]利用模型最后一个隐藏层子空间的核密度估计和贝叶斯神经网络不确定性估计,并结合逻辑回归模型检测输入是否为对抗样本,目前该研究只针对卷积神经网络(convolutional neural networks, CNN).Hendrycks 等人[117]利用对抗样本和真实样本之间的差异性提出了 3 种更加通用的检测方案,但这些方案共同的缺点是鲁棒性较差.为了避免数据处理影响真实样本的预测结果,有研究者提出只针对检测出对抗性的样本进行对抗噪声擦除操作.Meng 等人[118]利用一个检测网络来检测对抗样本,并利用一个重构网络重构对抗样本来消除其对抗性,结果显示该方案对黑盒攻击和灰盒攻击有较好的抵御效果.Cao 等人[119]提出了一种绕过对抗样本的方法.他们观察到对抗样本接近于分类边界,从输入空间中选择和此次预测样本接近的多个样本点进行预测,再由这些

样本点的预测结果进行投票作为此次预测结果.显然,该方案依赖于分类边界和样本点的选取.可见,现有对抗样本检测技术未充分考虑对抗样本生成机理,对攻击方法的先验知识和特定模型存在一定依赖性,检测能力有限.对抗性噪声擦除的研究以破坏对抗性噪声为主,但对抗性擦除操作会引入新的输入噪声,进而造成样本分类错误.

5 对抗性攻击与防御框架

从系统的角度出发,针对深度学习模型的攻击与防御不能是单一、片面和碎片化的,需要从完整流程、不同角度和层次上研究攻击与防御框架,才能真正有效地提高深度学习模型的安全性.

5.1 对抗性攻击框架

从对抗性攻击框架研究来看,现有工作[61,69,120-129]提出的对抗性攻击框架,只针对对抗样本生成,并且局限于特定模型与特定应用领域.Chang 等人[120]提出的对抗性攻击框架针对特殊的图嵌入模型生成对抗样本,并不是针对全周期深度学习模型的攻击面构建统一框架,方案局限性大、通用性低.Spampinato 等人[121]提出一个基于生成对抗网络的对抗性攻击框架,该框架通过自我监督机制学习视频表示动量特征,以便在视频中执行密集的全局预测,然而该框架也只局限于视频对抗样本生成.Chen 等人[122]基于 Frank-Wolfe 算法提出白盒和黑盒对抗样本对抗性攻击框架.该框架包含一个迭代的一阶白盒攻击算法以及带有 2 个感应矢量零阶优化选项的黑盒攻击算法,攻击手段同样局限于对抗样本.因此,以上研究只是将对抗样本的生成过程框架化,指导某些特定对抗样本生成,缺乏对抗性攻击的整体框架设计.

陈宇飞等人[130]针对人工智能系统安全与隐私风险,从人工智能系统基本框架的 4 个关键环节(输入、数据预处理、机器学习模型、输出)出发,指出相应的安全风险和应对措施,但是该工作仅仅针对现有技术进行了讨论和分类,并未形成一个细致完备的攻击或防御框架.

研究者们还实现了一些关于对抗性攻击框架的工具.2017 年德国图宾根大学的 3 名研究人员创建了一个基于 Python 库的 Foolbox 攻击框架[131],提供超过 20 多种类型的对抗性攻击手段,允许用户自定义攻击目标(比如错误分类)进而发动相应地攻击,然而该框架只注重对抗样本攻击,且攻击目标仅限于现有的机器学习框架.Nicolae 等人[132]开发了 Adversarial Robustness Toolbox 工具箱,其中实施的攻击包括逃逸攻击、提取攻击和投毒攻击,并使用目前最先进的威胁模型测试系统防御能力,以此来提高样本对抗性.然而,该工具箱未考虑模型训练阶段的后门攻击,未形成多层次多角度的统一攻击框架.

在工业界,2018 年 360 安全研究院发布的《AI 安全风险白皮书》[133]结合深度学习逃逸攻击方面的实例和研究工作,详细解读了人工智能应用所面临的安全风险.2019 年在《人工智能安全标准化白皮书》[134]中,清华大学研发了包含系统层、算法层和应用层的 RealSafe 人工智能安全平台架构,其中系统层提供算法的共性模块,算法层实现不同攻击环境下的对抗样本攻击.然而,该系统架构只考虑了单纯的对抗样本攻击,并不是一个针对深度学习模型的统一模型框架.

因此,虽然对抗性攻击得到了广泛关注和研究,但是现有攻击方案通常只考虑了深度学习模型的单个或部分攻击面,不同攻击模块之间缺乏联系,未形成统一的模型框架.

5.2 对抗攻击的防御框架

目前,也有不少研究工作针对已有攻击手段提出了相应的防御框架[81,111,135-143],这些防御框架包括数据异常检测框架[81]、模型增强框架[135-137]、对抗样本检测框架[111,138-139]、对抗噪声擦除框架[140-141]和隐私保护框架[142-143],涉及从数据收集到模型训练再到模型线上部署的每个阶段,包含模型鲁棒性和模型机密性等内容.然而,这些工作仅是框架化了单个防御方案的设计,不同防御策略相互之间缺乏协同,难以形成统一防御框架.

少数研究工作开始探索并建立了不同防御方案之间的联系.Wang 等人[144]提出模型诊断与模型增强的联合框架,首先分析网络每一层潜在脆弱性,然后提出相应的策略来优化训练过程以应对这些脆弱性.Akhtar 等人[145]将对抗样本检测和对抗噪声消除结合,通过一个输入重整网络对输入进行处理,并比较处理前后样本之间的差异性来确认输入样本是否包含对抗性扰动,从而决定输入到目标模型中的样本.尽管联合考虑了部分防御方案,这些框架仍然无法覆盖深度学习模型的完整生命周期.

在工业界,部分企业也提出了人工智能系统安全防御框架.在《人工智能安全标准化白皮书》[134]中,依托网络安全和安全管理将人工智能系统分为

云侧、边缘侧及端侧 3 个部分,建立了人工智能系统的安全框架体系,从宏观上确立人工智能系统的安全内涵,但未涉及具体的防御策略.2019 年华为技术有限公司发布了《AI 安全白皮书》[146],从攻防安全、模型安全及架构安全 3 个层次确立了 AI 安全防御架构,并针对每种潜在威胁提出了相应的防御手段,但是同样缺乏对防御手段之间关联性的考虑,无法建立有效的协同防御手段.

6 深度学习模型攻防框架与未来挑战

本文围绕对抗性攻击生成机理、对抗性攻击生成方法、对抗攻击的防御策略以及对抗攻击与防御框架 4 个方面对现有工作进行了阐述,并总结了相关工作的不足.从目前已有研究来看,我们认为今后针对深度学习模型攻击与防御的研究,应从“对抗”的角度出发,以攻击能力增强促进防御能力提升,着力研究针对深度学习模型的攻防分析基础理论、统一框架与关键技术.在本节,我们首先在现有工作的基础上分别提出针对深度学习模型的攻击与防御框架,然后总结现有挑战并展望未来研究方向,以期引起相关研究者的关注并提供指导.

6.1 针对深度学习模型的攻击框架

现有针对深度学习模型对抗性攻击生成方法的研究存在通用性差、假设条件强等问题.如何充分利用深度学习模型训练和推理阶段的固有缺陷,在保证攻击成功率的前提下设计更加实用、高效的对抗性攻击是尚待解决的难题.在攻击框架方面,现有的框架只是将攻击样本的生成过程框架化,指导攻击样本生成,仍是只考虑了模型的单个或部分攻击面,不同的攻击方法之间没有联动.本文从深度学习的整个生命周期角度出发,分析模型的脆弱性和攻击方式并提出了如图 2 所示的针对深度学习模型的攻击框架:

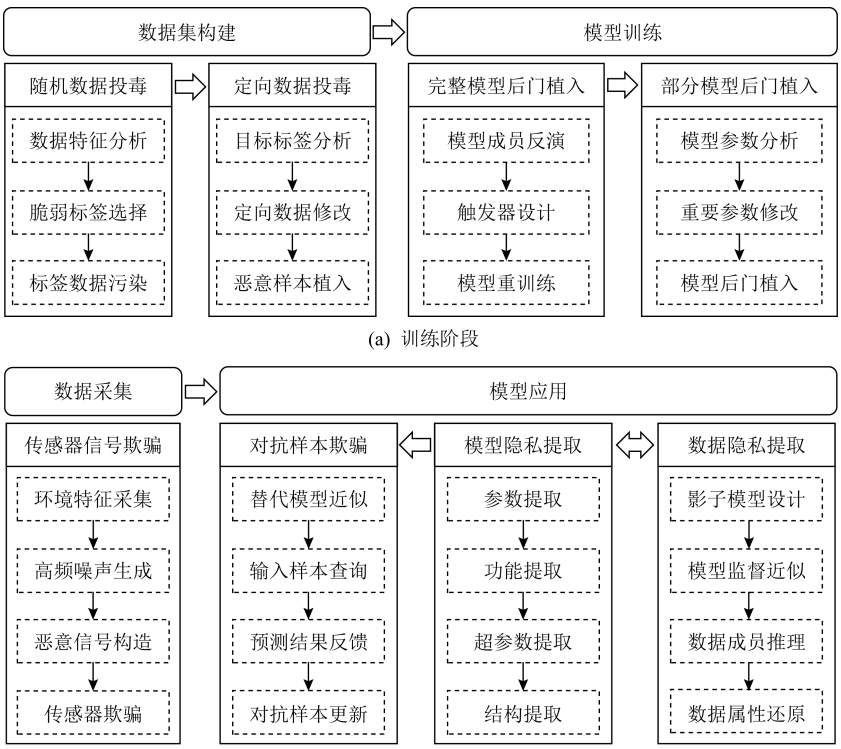


Fig. 2 Attack framework for deep learning models

图 2 针对深度学习模型的攻击框架

首先在训练阶段构建训练数据集时,可以发起随机数据投毒攻击和定向数据投毒攻击,训练模型过程中进一步完成完整模型后门植入攻击和部分模型后门植入攻击.然后在推理阶段获取输入样本时进行传感器信号欺骗攻击,在模型应用时进行对抗本欺骗攻击,在模型应用时实现模型和数据隐私提取攻击.训练阶段和推理阶段的攻击共同实现了深度学习模型全生命周期的攻击.

同时,为了实现更加有效的攻击,在我们提出的框架中各个模块之间进行可以相互协作.在训练

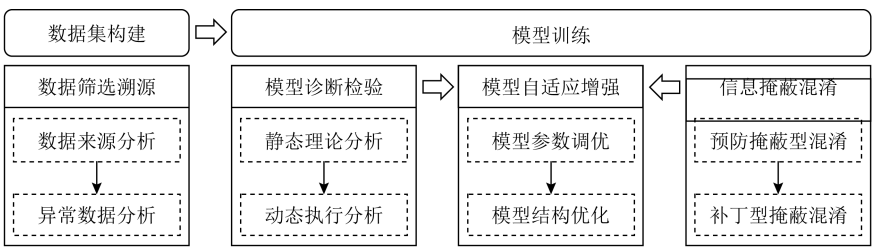
阶段,随机数据投毒可以为定向数据投毒提供支撑,完整模型后门植入可以为部分模型后门植入提供支撑.数据投毒攻击的结果也可以进一步支持后门植入攻击,针对完整模型的后门植入方案可以利用数据投毒的方式进行.在推理阶段,同时模型隐私提取的结果也可以进一步用来提高对抗样本欺骗的攻击能力,黑盒环境下可以利用模型提取的结果构造对抗样本攻击的替代模型,利用替代模型生成对抗样本进而攻击黑盒目标模型.同时训练和推理也可以相互协作,考虑到带有错误标签的投毒数据容易被发现,可以利用对抗样本生成算法生成投毒数据,实现高隐蔽的数据投毒攻击.

总的来说,我们构建的对抗性攻击的统一模型框架综合考虑了模型训练和推理阶段的脆弱性,将多个攻击面结合起来,联动各个攻击模块,充分利用不同攻击方法的优势,相互进行信息补充以弱化或消除阻碍攻击实用性的假设条件,模块之间的相互协作也进一步帮助实现更加全面有效的攻击.

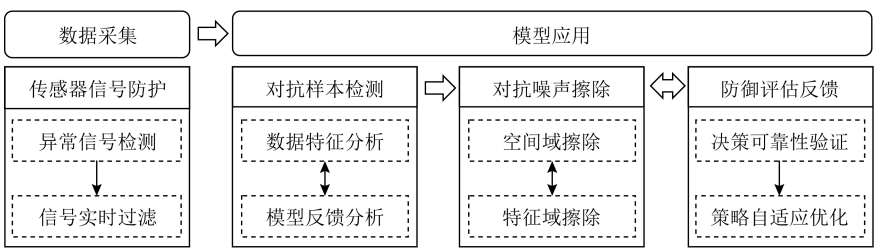
本文提出的对抗性攻击统一模型框架和生成方法可为不同应用领域不同深度学习模型的对抗性攻击提供指导,同时针对特定场景限制生成有针对性的攻击.传感器信号欺骗、数据投毒、后门植入、对抗样本欺骗、模型和数据隐私提取技术均具有一定的普适性.

6.2 针对深度学习模型的防御框架

在防御框架方面,现有方法只联合考虑了部分防御方案,但这些方法只是针对某一种攻击来设计,仍没有覆盖深度学习模型的完整生命周期,因此无法提供全面的防御能力.本文提出如图 3 所示的针对深度学习模型的防御框架,联合多种防御方案,以保证模型可用性为前提,充分考虑各防御方案的特点,从数据、算法等多个维度构建针对训练阶段、推理阶段的全方位防御框架:



(a) 线下防御



(b) 线上防御

Fig. 3 Defense framework for deep learning models

图 3 针对深度学习模型的防御框架

“线下防御”是在模型发布前从数据集构建和模型训练 2 个角度出发,在建立可靠数据集的基础上进一步提升模型的鲁棒性和机密性,最大化模型的防御能力.“线上防御”是在模型发布之后从输入样本获取和模型应用 2 个角度出发,在确保传感器数据准确的基础上,实时检测和擦除输入样本中的对抗噪声,并通过建立有效的评估反馈机制实现防御的动态优化.

在线下防御阶段,结合数据溯源技术和异常检测技术设计数据筛查溯源方案,以实现对数据的有效监管,确保训练集的可靠性.在此基础上,利用静态理论分析的方法对训练出的模型的脆弱性进行分析,并利用模型增强方法对模型参数进行调整以修复其脆弱性.为进一步增强模型的鲁棒性,对多个鲁棒模型进行安全聚合,并通过动态执行分析对模型聚合策略进行调整,提升模型的聚合效率.与此同时,针对模型发布前的各个阶段,设计满足多种隐私保护需求的信息混淆掩蔽方案,对数据集和模型参数进行有效保护.线下协同防御策略正是联合防御方案,从可靠性、鲁棒性以及机密性 3 个角度全面提升模型的防御能力:

在线上防御阶段,可以通过异常信号检测和信号实时过滤的方法对输入信号中的异常频段信号进行实时的检测过滤,确保传感器数据符合深度学习

模型的需求.同时,为进一步确保输入样本中不包含对抗噪声,在对抗样本检测之后,进一步执行对抗噪声擦除操作,对抗噪声擦除的结果可以反馈给对抗样本检测步骤对检测方法进行动态调整,同时也可通过对抗样本检测来验证对抗噪声擦除的效果.除此之外,还应该对输入样本的预测结果进行分析,评估此次预测结果的可靠性,并根据评估结果调整深度学习模型的防御方案,实现模型防御动态优化.线上协同防御策略通过联合检测、擦除防御方案,动态调整防御策略,对执行预测任务的输入样本进行多层次的保护,确保预测结果的准确性.

同时线上防御阶段针对对抗样本的检测结果可以反馈给线下防御的用于加固深度学习模型,对模型增强方法和防御策略进行调整,提升深度学习模型的鲁棒性.面对未知攻击,可以持续动态评估模型在现有环境下的脆弱性,并根据分析结果实时调整多模型的聚合方式,以增强模型对新型攻击的抵抗能力.

对抗攻击的统一防御框架及协同防御可为不同应用领域不同深度学习模型的防御策略设计提供通用性指导,同时针对常见攻击实施特定防御.数据溯源、模型增强和信息掩蔽混淆、传感器信号防护、对抗样本检测与对抗噪声擦除、防御评估反馈技术均具有一定的普适性.

6.3 深度学习模型攻击与防御研究的未来挑战

目前,深度学习模型对抗攻击与防御研究面临的问题与挑战主要集中在 3 个方面:

1) 难以针对深度学习模型的对抗性攻击本质建立通用、完整的机理分析体系.深度学习模型的对抗性攻击的本质是利用模型固有缺陷,构造恶意输入进而导致模型行为异常.因此,研究对抗性攻击机理是分析弥补模型内部缺陷、探索高效攻防技术的基础支撑.然而,深度学习数据规模庞大、模型结构复杂、可解释性差,目前关于对抗性攻击生成机理的研究解释还不够充分,尚未有更加系统、通用的理论分析手段,造成现有绝大多数对抗性攻击生成或防御方法具有很大的局限性.因此,如何多角度、全方位地分析深度学习模型各阶段对抗性攻击的生成机理,进而发现其背后更深层次的通用性原理,指导设计更为高效的攻击和防御方法,是未来研究要解决的关键问题之一.

2) 面向非确定目标时的对抗性攻防难以有效实施.现有针对深度学习模型的攻击和防御方案大都依赖于对目标模型和目标攻击的先验知识,然而实际场景中存在目标模型隐私信息获取困难、各种攻击手段层出不穷的问题,削弱了攻击和防御方案的实际效果.以对抗样本生成攻击为例,首先,在严格黑盒的条件下,攻击者只能访问模型返回结果的硬标签,这给对抗样本生成带来了巨大的挑战;其次,如何在真实物理情况下发起攻击、抵抗环境噪声干扰,也是待解决的一大难题;最后,如何减少访问次数,同时防止被攻击方的察觉,也是在该领域极具挑战的问题之一.

3) 难以构建全面覆盖各攻击面的攻击和防御框架.现有针对深度学习模型的对抗性攻击和防御策略设计通常是孤立分散的,即攻击和防御只针对模型生命周期某一阶段的单个攻击面来制定,缺乏统一的攻击和防御指导框架.单一的攻击和防御方案在强博弈对抗的动态环境下难以达到预期效果.这导致攻击难以攻其无备,防御也防不胜防.简单地组合攻击方案,可能会降低攻击行为隐蔽性,同时加大攻击成本,削弱攻击效率.而单纯地集成多种防御方案也难以提升模型防御能力,反而可能会降低模型的可用性.所以,如何设计全覆盖的协同攻击和协同防御框架是未来研究需要攻克的问题之一.

7 结束语

人工智能研究的快速发展和深度学习技术在实际场景中的广泛应用吸引了一大批来自于学术界和工业界的学者的深入研究,并取得了令人瞩目的研究成果.然而,深度学习技术在安全性上还存在诸多问题,其潜在的安全隐患和隐私风险为人工智能的全面部署带来挑战.为了重新审视深度学习模型攻击与防御的研究现状,梳理现有研究成果的优势与不足,明确未来研究方向,本文系统地从对抗性攻击生成机理、对抗性攻击生成方法、对抗攻击的防御策略、对抗性攻击与防御框架 4 个方面研究了深度学习模型安全性问题,回顾了大量的极具影响力的研究成果并对相关研究进行了科学的总结和归纳.最后,本文指出了深度学习模型隐私保护研究当前面临的挑战,并探讨了未来可行的研究方向,旨在推动针对深度学习模型对抗性攻防研究的进一步发展.

参 考 文 献

[1] Kloft M, Laskov P. Online anomaly detection under adversarial impact [C] //Proc of the 13th Artificial Intelligence and Statistics (AISTATS). Cambridge, MA: JMLR, 2010: 405-412

[2] 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(AAAI). Menlo Park, CA: AAAI, 2015: 2871-2877

[3] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with backgradient optimization [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York: ACM, 2017: 27-38

[4] Biggio B, Fumera G, Roli F. Design of robust classifiers for adversarial environments [C] //Proc of IEEE Int Conf on Systems, Man, and Cybernetics (SMC). Piscataway, NJ: IEEE, 2011: 977-982

[5] Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning [C] //Proc of the 39th IEEE Symp on Security and Privacy(S&P). Piscataway, NJ: IEEE, 2018: 19-35

[6] Chen Yudong, Caramanis C, Mannor S, et al. Robust high dimensional sparse regression and matching pursuit [J]. arXiv preprint, arXiv:1301.2725, 2013

[7] Liu Chang, Li Bo, Vorobeychik Y, et al. Robust linear regression against training data poisoning [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York: ACM, 2017: 91-102

[8] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks [C] //Proc of the 25th Network and Distributed System Security Symp (NDSS). Reston, VA: ISOC, 2018

[9] Ji Yujie, Zhang Xinyang, Wang Ting. Backdoor attacks against learning systems [C] //Proc of the 5th IEEE Conf on Communications and Network Security (CNS). Piscataway, NJ: IEEE, 2017

[10] McDaniel P, Papernot N, Celik Z B. Machine learning in adversarial settings [C] //Proc of the 37th IEEE Symp on Security and Privacy(S&P). Piscataway, NJ: IEEE, 2016: 68-72

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C] //Proc of the 2nd Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2014

[12] Pei Kexin, Cao Yinzhi, Yang Junfeng, et al. Deepxplore: Automated whitebox testing of deep learning systems [C] //Proc of the 26th ACM Symp on Operating Systems Principles (SOSP). New York: ACM, 2017: 1-18

[13] Song Yang, Kim T, Nowozin S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples [C] //Proc of the 6th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2018

[14] Gilmer J, Ford N, Carlini N, et al. Adversarial examples are a natural consequence of test error in noise [C] //Proc of the 36th Int Conf on Machine Learning (ICML). New York: ACM, 2019: 2280-2289

[15] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially robust generalization requires more data [C] //Proc of the 32nd Annual Conf on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2018: 5019-5031

[16] Gilmer J, Metz L, Faghri F, et al. Adversarial spheres [C] //Proc of the 6th Int Conf on Learning Representations Workshop Track(ICLRW). La Jolla, CA: ICLR, 2018

[17] Mahloujifar S, Diochnos D I, Mahmoody M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence(AAAI). Menlo Park, CA: AAAI, 2019: 4536-4543

[18] Shafahi A, Huang W R, Studer C, et al. Are adversarial examples inevitable? [C] //Proc of the 7th Int Conf on Learning Representations(ICLR). La Jolla, CA: ICLR, 2019

[19] Bubeck S, Price E, Razenshteyn I. Adversarial examples from computational constraints [C] //Proc of the 36th Int Conf on Machine Learning(ICML). New York: ACM, 2019: 831-840

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [C] //Proc of the 36th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2015

[21] Tanay T, Griffin L. A boundary tilting persepective on the phenomenon of adversarial examples [J]. arXiv preprint, arXiv:1608.07690, 2016

[22] Fawzi A, Fawzi O, Frossard P. Fundamental limits on adversarial robustness [C] //Proc of the 32nd Int Conf on Machine Learning (ICML), Workshop on Deep Learning. New York: ACM, 2015

[23] Papernot N, McDaniel P, Goodfellow I J. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples [J]. arXiv preprint, arXiv: 1605.07277, 2016

[24] Tabacof P, Valle E. Exploring the space of adversarial images [C] //Proc of Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2016: 426-433

[25] He W, Li Bo, Song D. Decision boundary analysis of adversarial examples [C] //Proc of the 6th Int Conf on Learning Representations(ICLR). La Jolla, CA: ICLR, 2018

[26] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features [C] //Proc of the 33rd Annual Conf on Neural Information Processing Systems(NeurIPS). Cambridge, MA: MIT Press, 2019: 125-136

[27] Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture: increasing shape bias improves accuracy and robustness [C] //Proc of the 7th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2019

[28] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 3681-3688

[29] Nar K, Ocal O, Sastry S S, et al. Cross-entropy loss and low-rank featureshave responsibility for adversarial examples [J]. arXiv preprint, arXiv:1901.08360, 2019

[30] Pang Tianyu, Xu Kun, Dong Yinpeng, et al. Rethinking softmax cross-entropy loss for adversarial robustness [C] //Proc of the 7th Int Conf on Learning Representations(ICLR). La Jolla, CA: ICLR, 2019

[31] Bartlett P L, Foster D J, Telgarsky M J. Spectrally-normalized margin bounds for neural networks [C] //Proc of the 31st Annual Conf on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2017: 6240-6249

[32] Zhang Chiyuan, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization [C] //Proc of the 5th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2017

[33] Fabio C, Rudy B, Roberto C, et al. Adversarial examples detection in features distance spaces [C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 313-327

[34] Noack A, Ahern I, Dou Dejing, et al. Does interpretability of neural networks imply adversarial robustness [J]. arXiv preprint, arXiv:1912.03430, 2019

[35] Dong Yinpeng, Su Hang, Zhu Jun, et al. Towards interpretable deep neural networks by leveraging adversarial examples [J]. arXiv preprint, arXiv:1708.05493, 2017

[36] Paudice A, Muñoz-González L, Gyorgy A, et al. Detection of adversarial training examples in poisoning attacks through anomaly detection [J]. arXiv preprint, arXiv:1802.03041, 2018

[37] Grosse K, Manoharan P, Papernot N, et al. On the (statistical) detection of adversarial examples [J]. arXiv preprint, arXiv:1702.06280, 2017

[38] Wei Fan, Song Yunfei, Shao Mingli, et al. Improving adversarial robustness on single model via feature fusion and ensemble diversity [J]. Journal of Software, 2020, 31(9): 2756-2769 (in Chinese)
(韦璠, 宋云飞, 邵明莉, 等. 利用特征融合和整体多样性提升单模型鲁棒性[J]. 软件学报, 2020, 31(9): 2756-2769)

[39] Theagarajan R, Chen Ming, Bhanu B, et al. ShieldNets: Defending against adversarial attacks using probabilistic adversarial robustness [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 6988-6996

[40] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York: ACM, 2017: 3-14

[41] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 38th IEEE Symp on Security and Privacy(S&P). Piscataway, NJ: IEEE, 2017: 39-57

[42] Kearnsl M, Li Ming. Learning in the presence of malicious errors [J]. SIAM Journal on Computing, 1993, 22(4): 807-837

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

[44] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines [C] //Proc of the 29th Int Conf on Machine Learning(ICML). New York: ACM, 2012

[45] Gu Tianyu, Liu Kang, Dolan-Gavitt B, et al. BadNets: Evaluating backdooring attacks on deep neural networks [J]. IEEE Access, 2019, 7: 47230-47244

[46] Ji Yujie, Zhang Xinyang, Ji Shouling, et al. Model-reuse attacks on deep learning systems [C] //Proc of the 25th ACM Conf on Computer and Communications Security(CCS). New York: ACM, 2018: 349-363

[47] Chen Xinyun, Liu Chang, Li Bo, et al. Targeted backdoor attacks on deep learning systems using data poisoning [J]. arXiv preprint, arXiv:1712.05526, 2017

[48] Tramèr F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction apis [C] //Proc of the 25th USENIX Security Symp(USENIX Security). Berkeley, CA: USENIX Association, 2016: 601-618

[49] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 4954-4963

[50] Juuti M, Szyller S, Marchal S, et al. Prada: Protecting against DNN model stealing attacks [C] //Proc of the 4th IEEE European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2019: 512-527

[51] Duddu V, Samanta D, Rao D V, et al. Stealing neural networks via timing side channels [J]. arXiv preprint, arXiv: 1812.11720, 2018

[52] Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays [J]. PLoS Genetics, 2008, 4(8): No.e1000167

[53] Backes M, Berrang P, Humbert M, et al. Membership privacy in MicroRNA-based studies [C] //Proc of the 32nd ACM Conf on Computer and Communications Security (CCS). New York: ACM, 2016: 319-330

[54] Dwork C, Smith A, Steinke T, et al. Robust traceability from trace amounts [C] //Proc of the 56th IEEE Annual Symp on Foundations of Computer Science (FOCS). Piscataway, NJ: IEEE, 2015: 650-669

[55] Calandrino J A, Kilzer A, Narayanan A, et al. You might also like: Privacy risks of collaborative filtering [C] //Proc of the 32nd IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2011: 231-246

[56] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing [C] //Proc of the 23rd USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2014: 17-32

[57] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 22nd ACM Conf on Computer and Communications Security (CCS). New York: ACM, 2015: 1322-1333

[58] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2015: 427-436

[59] Papernot N, McDaniel P, Goodfellow I J, et al. Practical black-box attacks against machine learning [C] //Proc of the 12th ACM Asia Conf on Computer and Communications Security (AsiaCCS). New York: ACM, 2017: 506-519

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

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

[62] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 2574-2582

[63] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 1765-1773

[64] Xiao Chaowei, Li Bo, Zhu J Y, et al. Generating adversarial examples with adversarial networks [J]. arXiv preprint, arXiv:1801.02610, 2018

[65] Su Jiawei, Vargas D V, Kouichi S. One-pixel attack for fooling deep neural networks [J]. arXiv preprint, arXiv: 1710.08864, 2017

[66] Liu Yanpei, Chen Xinyun, Liu Chang, et al. Delving into transferable adversarial examples and black-box attacks [C] //Proc of the 5th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2017

[67] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 9185-9193

[68] Chen P Y, Zhang Huan, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York: ACM, 2017: 15-26

[69] Tu C C, Ting Paishun, Chen P Y, et al. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 742-749

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

[71] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models [C] //Proc of the 6th Int Conf on Learning Representations (ICLR). La Jolla, CA: ICLR, 2018

[72] Thomas A H, Bhavya K. Universal hard-label black-box perturbations: Breaking security-through-obscurity defenses [J]. arXiv preprint, arXiv:1811.03733, 2018

[73] Chen Jinyin, Shen Shijing, Su Mengmeng, et al. Black-box Adversarial Attack on License Plate Recognition System [J]. Acta Automatica Sinica, 2021, 47(1): 121-135 (in Chinese) (陈晋音, 沈诗婧, 苏蒙蒙, 等. 车牌识别系统的黑盒对抗攻击[J]. 自动化学报, 2021, 47(1): 121-135

[74] Hu Weiwei, Tan Ying. Generating adversarial malware examples for black-box attacks based on GAN [J]. arXiv preprint, arXiv: 1702. 05983, 2017

[75] Sarkar S, Bansal A, Mahbub U, et al. UPSET and ANGRI: Breaking high performance image classifiers [J]. arXiv preprint, arXiv:1707.01159, 2017

[76] Yoo Y, Park S, Choi J, et al. Butterfly effect: Bidirectional control of classification performance by small additive perturbation [J]. arXiv preprint, arXiv:1711.09681, 2017

[77] Zhao Zhengli, Dua D, Singh S. Generating natural adversarial examples [J]. arXiv preprint, arXiv:1710.11342, 2017

[78] Baracaldo N, Chen B, Ludwig H, et al. Detecting poisoning attacks on machine learning in iot environments [C] //Proc of the 2nd IEEE Int Congress on Internet of Things (ICIOT). Piscataway, NJ: IEEE, 2018: 57-64

[79] Baracaldo N, Chen B, Ludwig H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York: ACM, 2017: 103-110

[80] Liu Yuntao, Xie Yang, Srivastava A. Neural trojans [C] //Proc of the 35th IEEE Int Conf on Computer Design (ICCD). Piscataway, NJ: IEEE, 2017: 45-48

[81] Steinhardt J, Koh P W W, Liang P S. Certified defenses for data poisoning attacks [C] //Proc of the 31st Annual Conf on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2017: 3517-3529

[82] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering [J]. arXiv preprint, arXiv:1811.03728, 2018

[83] Paudice A, Muñoz-González L, Lupu E C. Label sanitization against label flipping poisoning attacks [C] //Proc of Joint European Conf on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Berlin: Springer, 2018: 5-15

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

[85] Kurakin A, Boneh D, Tramèr F, et al. Ensemble adversarial training: Attacks and defenses [J]. arXivpreprint, arXiv: 1705.07204, 2017

[86] Gu Shixiang, Rigazio L. Towards deep neural network architectures robust to adversarial examples [J]. arXivpreprint, arXiv:1412.5068, 2014

[87] Lee H, Han S, Lee J. Generative adversarial trainer: Defense to adversarial perturbations with gan [J]. arXivpreprint, arXiv:1705.03387, 2017

[88] Bradshaw J, Matthews A G D G, Ghahramani Z. Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks [J]. arXiv preprint, arXiv:1707.02476, 2017

[89] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] //Proc of the 37th IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2016: 582-597

[90] Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096)

[91] Gopinath D, Katz G, Pasareanu C S, et al. Deepsafe: A data-driven approach for checking adversarial robustness in neural networks [J]. arXivpreprint, arXiv:1710.00486, 2017

[92] Koh P W, Liang P. Understanding black-box predictions via influence functions [C] //Proc of the 34th Int Conf on Machine Learning (ICML). New York: ACM, 2017: 1885-1894

[93] Graepel T, Lauter K, Naehrig M. ML confidential: Machine learning on encrypted data [C] //Proc of the 15th Annual Int Conf on Information Security and Cryptology (ICISC). Berlin: Springer, 2012: 1-21

[94] Abadi M, Chu A, Goodfellow I J, et al. Deep learning with differential privacy [C] //Proc of the 23rd ACM Conf on Computer and Communications Security (CCS). New York: ACM, 2016: 308-318

[95] Zhang Jiaqi, Zheng Kai, Mou Wenlong, et al. Efficient private ERM for smooth objectives [J]. arXivpreprint, arXiv:1703.09947, 2017

[96] Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization [C] // Proc of the 25th ACM Conf on Computer and Communications Security (CCS). New York: ACM, 2018: 634-646

[97] Tripathy A, Wang Ye, Ishwar P. Privacy-preserving adversarial networks [C] //Proc of the 57th Annual Allerton Conf on Communication, Control, and Computing (Allerton). Piscataway, NJ: IEEE, 2019: 495-505

[98] Hesamifard E, Takabi H, Ghasemi M. Cryptodl: Deep neural networks over encrypted data [J]. arXivpreprint, arXiv:1711.05189, 2017

[99] Aono Y, Hayashi T, Wang Lihua, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345

[100] Tramèr F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction apis [C] //Proc of the 25th USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2016: 601-618

[101] Wang Binghui, Gong N Z. Stealing hyperparameters in machine learning [C] //Proc of the 39th IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2018: 36-52

[102] Selvaraj J, Dayanıklı G Y, Gaunkar N P, et al. Electromagnetic induction attacks against embedded systems [C] //Proc of the 13th ACM Asia Conf on Computer and Communications Security (AsiaCCS). New York: ACM, 2018: 499-510

[103] Giechaskiel I, Zhang Y, Rasmussen K B. A framework for evaluating security in the presence of signal injection attacks [C] //Proc of the 24th European Symp on Research in Computer Security (ESORICS). Berlin: Springer, 2019: 512-532

[104] Tu Yazhou, Lin Zhiqiang, Lee I, et al. Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors [C] //Proc of the 27th USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2018: 1545-1562

[105] Khazaaleh S, Korres G, Eid M, et al. Vulnerability of MEMS gyroscopes to targeted acoustic attacks [J]. IEEE Access, 2019, 7: 89534-89543

[106] Mahmoud D, Stojilović M. Timing violation induced faults in multi-tenant FPGAs [C] //Proc of the 23rd Design, Automation and Test in Europe Conf and Exhibition (DATE). Piscataway, NJ: IEEE, 2019: 1745-1750

[107] Muniraj D, Farhood M. Detection and mitigation of actuator attacks on small unmanned aircraft systems [J]. Control Engineering Practice, 2019, 83: 188-202

[108] Kune D F, Backes J, Clark S S, et al. Ghost talk: Mitigating EMI signal injection attacks against analog sensors [C] //Proc of the 34th IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2013: 145-159

[109] Nashimoto S, Suzuki D, Sugawara T, et al. Sensor CONFusion: Defeating kalman filter in signal injection attack [C] //Proc of the 13th ACM Asia Conf on Computer and Communications Security (AsiaCCS). New York: ACM, 2018: 511-524

[110] Wang Jingyi, Sun Jun, Zhang Peixin, et al. Detecting adversarial samples for deep neural networks through mutation testing [J]. arXivpreprint, arXiv: 1805.05010, 2018

[111] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [J]. arXivpreprint, arXiv:1704.01155, 2017

[112] Das N, Shanbhogue M, Chen S T, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression [J]. arXivpreprint, arXiv:1705.02900, 2017

[113] Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of jpg compression on adversarial images [J]. arXivpreprint, arXiv:1608.00853, 2016

[114] Xie Cihang, Wang Jianyu, Zhang Zhishuai, et al. Mitigating adversarial effects through randomization [J]. arXivpreprint, arXiv:1711.01991, 2017

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

[116] Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts [J]. arXivpreprint, arXiv:1703.00410, 2017

[117] Hendrycks D, Gimpel K. Early methods for detecting adversarial images [J]. arXivpreprint, arXiv: 1608.00530, 2016

[118] Meng Dongyu, Chen Hao. Magnet: A two-pronged defense against adversarial examples [C] //Proc of the 24th ACM Conf on Computer and Communications Security (CCS). New York: ACM, 2017: 135-147

[119] 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 (ACSAC). Piscataway, NJ: IEEE, 2017: 278-287

[120] Chang Heng, Rong Yu, Xu Tingyang, et al. A restricted black-box adversarial framework towards attacking graph embedding models [C] //Proc of the 34th AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2020: 3389-3396

[121] Spampinato C, Palazzo S, D'Oro P, et al. Adversarial framework for unsupervised learning of motion dynamics in videos [J]. International Journal of Computer Vision, 2020, 128(5): 1378-1397

[122] Chen Jinghui, Zhou Dongruo, Yi Jinfeng, et al. A Frank-Wolfe framework for efficient and effective adversarial attacks [J]. arXivpreprint, arXiv:1811.10828, 2018

[123] Jin Di, Jin Zhijing, Zhou J T, et al. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment [C] //Proc of the 34th AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2020: 8018-8025

[124] Zhao Pu, Liu Sijia, Wang Yanzhi, et al. An admm-based universal framework for adversarial attacks on deep neural networks [C] //Proc of the 26th ACM Int Conf on Multimedia (ACM Multimedia). New York: ACM, 2018: 1065-1073

[125] Zheng Tianhang, Chen Changyou, Ren Kui. Distributionally adversarial attack [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 2253-2260

[126] Behzadan V, Munir A. Models and framework for adversarial attacks on complex adaptive systems [J]. arXiv preprint, arXiv:1709.04137, 2017

[127] Yu Fuxun, Dong Qide, Chen Xiang. Asp: A fast adversarial attack example generation framework based on adversarial saliency prediction [J]. arXivpreprint, arXiv: 1802.05763, 2018

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

[129] Behzadan V, Munir A. Adversarial reinforcement learning framework for benchmarking collision avoidance mechanisms in autonomous vehicles [J/OL]. IEEE Intelligent Transportation Systems Magazine, 2019 [2020-09-12]. https://ieeexplore.ieee.org/abstract/document/8686215

[130] Chen Yufei, Shen Chao, Wang Qian, et al. Security and privacy risks in artificial intelligence systems [J]. Journal of Computer Research and Development, 2019, 56 (10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险 [J]. 计算机研究与发展, 2019, 56(10): 2135-2150)

[131] Rauber J, Brendel W, Bethge M. Foolbox: A Python toolbox to benchmark the robustness of machine learning models [J]. arXiv preprint, arXiv:1707.04131, 2017

[132] Nicolae M I, Sinn M, Tran M N, et al. Adversarial Robustness Toolbox v1.0.0 [J]. arXivpreprint, arXiv:1807.01069, 2018

[133] 360 Security Research Institute. AI Security Risk White Paper [EB/OL]. [2020-09-12]. https://www.freebuf.com/articles/network/162875.html (in Chinese)
(360 安全研究院. AI 安全风险白皮书[EB/OL]. [2020-09-12]. https://www.freebuf.com/articles/network/162875.html)

[134] National Information Security Standardization Technical Committee. Artificial Intelligence Security Standardization White Paper [EB/OL]. [2020-09-12]. http://www.cesi.cn/201911/5733.html (in Chinese)
(全国信息安全标准化技术委员会. 人工智能安全标准化白皮书[EB/OL]. [2020-09-12]. http://www.cesi.cn/201911/5733.html)

[135] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization [J]. Neurocomputing, 2018, 307: 195-204

[136] Li Bo, Vorobeychik Y, Chen Xinyun. A general retraining framework for scalable adversarial classification [J]. arXiv preprint, arXiv:1604.02606, 2016

[137] Boopathy A, Weng T W, Chen P Y, et al. CNN-Cert: An efficient framework for certifying robustness of convolutional neural networks [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 3240-3247

[138] Liu Ninghao, Yang Hongxia, Hu Xia, et al. Adversarial detection with model interpretation [C] //Proc of the 24th ACM Knowledge Discovery and Data Mining (SIGKDD). New York: ACM, 2018: 1803-1811

[139] Ruan Yibin, Dai Jiazhu. TwinNet: A double sub-network framework for detecting universal adversarial perturbations [J]. Future Internet, 2018, 10(3): 26-26

[140] Shen Shiwei, Jin Guoqing, Gao Ke, et al. Ape-gan: Adversarial perturbation elimination with gan [J]. arXivpreprint, arXiv:1707.05474, 2017

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

[142] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning [C] //Proc of the 25th ACM Conf on Computer and Communications Security(CCS). New York: ACM, 2018: 35-52

[143] Juvekar C, Vaikuntanathan V, Chandrakasan A. GAZELLE: A low latency framework for secure neural network inference [C] //Proc of the 27th USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2018: 1651-1669

[144] Wang Shen, Chen Zhengzhang, Ni Jingchao, et al. Adversarial defense framework for graph neural network [J]. arXivpreprint, arXiv:1905.03679, 2019

[145] Akhtar N, Liu Jian, Mian A. Defense against universal adversarial perturbations [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 3389-3398

[146] Huawei Technologies Company Limited. AI Security White Paper [EB/OL]. [2020-09-12]. https://www-file.huawei.com/-/media/corporate/pdf/cyber-s ecurity/ai-security-white-paper-cn.pdf (in Chinese)
(华为技术有限公司. AI 安全白皮书[EB/OL]. [2020-09-12]. https://www-file.huawei.com/-/media/corporate/pdf/cyber-security/ai-security-white-paper-cn.pdf)

Li Minghui, born in 1994. PhD candidate. Her main research interests include security issues in cloud computing and artificial intelligence security.

李明慧,1994 年生.博士研究生.主要研究方向为云计算安全和人工智能安全.

Jiang Peipei, born in 1997. PhD candidate. Her main research interests include network security, information security, and applied cryptography.

江沛佩,1997 年生.博士研究生.主要研究方向为网络安全、信息安全和应用密码学.

Wang Qian, born in 1980. PhD, professor, PhD supervisor. His main research interests include AI security, data storage, search and computation outsourcing security and privacy protection, wireless systems security, big data security and privacy, and applied cryptography.

王 骞,1980 年生.博士,教授,博士生导师.主要研究方向为人工智能安全、数据存储、查询及计算外包安全与隐私保护、无线系统安全、大数据安全与隐私和应用密码学等.

Shen Chao, born in 1985. PhD, professor, PhD supervisor. His main research interests include cyber-physical system optimization and security, network and system security, and artificial intelligence security.

沈 超,1985 年生.博士,教授,博士生导师.主要研究方向为信息物理融合系统优化与安全、网络和系统安全和人工智能安全.

Li Qi, born in 1979. PhD, associate professor, PhD supervisor. His main research interests include network and system security, Internet security, mobile security, and big data security.

李 琦,1979 年生.博士,副教授,博士生导师.主要研究方向为网络和系统安全、互联网安全、移动安全和大数据安全等.