DOI:10.7544/issn1000-1239.2019.20190415
56(10):2135-2150,2019

人工智能系统安全与隐私风险

陈宇飞[1,2] 沈 超[1,2] 王 骞[3] 李 琦[4] 王 聪[5] 纪守领[6,7] 李 康[8] 管晓宏[1,2]

[1](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)

[2](西安交通大学电子与信息学部 西安 710049)

[3](武汉大学网络安全学院 武汉 430072)

[4](清华大学网络科学与网络空间研究院 北京 100084)

[5](香港城市大学计算机科学系 香港 999077)

[6](浙江大学网络空间安全研究中心 杭州 310027)

[7](浙江大学计算机科学与技术学院 杭州 310027)

[8](乔治亚大学计算机科学系 乔治亚州雅典市 30602)

(yfchen@sei.xjtu.edu.cn)

Security and Privacy Risks in Artificial Intelligence Systems

Chen Yufei[1,2], Shen Chao[1,2], Wang Qian[3], Li Qi[4], Wang Cong[5], Ji Shouling[6,7], Li Kang[8], and Guan Xiaohong[1,2]

[1](*Key Laboratory for Intelligent Networks and Network Security* (*Xi'an Jiaotong University*), *Ministry of Education*, *Xi'an* 710049)

[2](*Faculty of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)

[3](*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)

[4](*Institute for Network Sciences and Cyberspace*, *Tsinghua University*, *Beijing* 100084)

[5](*Department of Computer Science*, *City University of Hong Kong*, *Hong Kong* 999077)

[6](*Institute of Cyberspace Research*, *Zhejiang University*, *Hangzhou* 310027)

[7](*College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310027)

[8](*Department of Computer Science*, *University of Georgia*, *Athens*, *Georgia*, *the United States* 30602)

Abstract Human society is witnessing a wave of artificial intelligence (AI) driven by deep learning techniques, bringing a technological revolution for human production and life. In some specific fields, AI has achieved or even surpassed human-level performance. However, most previous machine learning theories have not considered the open and even adversarial environments, and the security and privacy issues are gradually rising. Besides of insecure code implementations, biased models, adversarial examples, sensor spoofing can also lead to security risks which are hard to be discovered by traditional security analysis tools. This paper reviews previous works on AI system security and privacy, revealing potential security and privacy risks. Firstly, we introduce a threat model of AI systems, including attack surfaces, attack capabilities and attack goals. Secondly, we analyze security risks and counter measures in terms of four critical components in AI systems: data input (sensor), data preprocessing, machine learning model and output. Finally, we discuss future research trends on the security of AI systems. The aim of this paper is to arise the attention of the computer security

收稿日期:2019-06-12;**修回日期**:2019-08-03

基金项目:国家自然科学基金项目(61822309,61773310,U1736205)

This work was supported by the National Natural Science Foundation of China (61822309, 61773310, U1736205).

通信作者:沈超(chaoshen@mail.xjtu.edu.cn)

society and the AI society on security and privacy of AI systems, and so that they can work together to unlock AI's potential to build a bright future.

Key words intelligent system security; system security; data processing; artificial intelligence (AI); deep learning

摘 要 人类正在经历着由深度学习技术推动的人工智能浪潮,它为人类生产和生活带来了巨大的技术革新.在某些特定领域中,人工智能已经表现出达到甚至超越人类的工作能力.然而,以往的机器学习理论大多没有考虑开放甚至对抗的系统运行环境,人工智能系统的安全和隐私问题正逐渐暴露出来.通过回顾人工智能系统安全方面的相关研究工作,揭示人工智能系统中潜藏的安全与隐私风险.首先介绍了包含攻击面、攻击能力和攻击目标的安全威胁模型.从人工智能系统的4个关键环节——数据输入(传感器)、数据预处理、机器学习模型和输出,分析了相应的安全隐私风险及对策.讨论了未来在人工智能系统安全研究方面的发展趋势.

关键词 智能系统安全;系统安全;数据处理;人工智能;深度学习

中图法分类号 TP391

近年来人工智能技术,尤其是深度学习理论方法,取得了重大突破.在计算机视觉[1]、语音识别[2-3]、自然语言处理[4]、棋牌博弈[5]等多类任务上,人工智能技术的判断准确水平和决策能力已经追平甚至超越人类.人工智能技术已经"走出实验室,跨入工业界"[6],迅速触及到人类生产和生活的方方面面.与此同时,人工智能技术开发日趋大众化.Caffe[7],Tensorflow[8],Torch[9],MXNet[10],PaddlePaddle[11]等开源深度学习框架提供了丰富的高级模块化函数支持,大大降低了应用的开发难度;腾讯[12]、阿里云[13]、百度[14]、谷歌[15]、微软[16]、亚马逊[17]、IBM[18]等厂商也都提供了人工智能服务,涵盖图像识别、语音识别、自动机器学习等多个方面.通过调用API接口,开发者可以实现高性能的人工智能应用.得益于理论与工具的发展,人工智能系统正大范围地部署.

然而,随着各类人工智能应用的出现和发展,其中的安全隐患也逐渐暴露出来.2018年3月发生在美国亚利桑那州的优步无人车事故中,事发时处于自动驾驶模式的无人车并没有检测到前方行人,驾驶员也未及时进行干预,最终致使行人被撞身亡[19].微软于2016年上线的社交机器人Tay,在一天之内受到用户的不良诱导逐渐学习成为一位种族主义者,迫使微软将该机器人紧急下线[20].在包括自动驾驶[21-22]、恶意软件检测[23]、视频安防[24]等在内的安全敏感领域,需对人工智能系统安全性和稳定性提出更高的要求.除了安全问题之外,隐私问题同样也受到人工智能服务提供商和用户的关注.由于机器模型的训练需要依赖大量的训练数据和计算资源,模型隐私与知识产权保护成为服务提供商最为关心的问题之一.而对用户而言,他们则更关注其个人信息作为训练数据是否会被泄露,如何才能确保个人敏感信息不被第三方窃取.

目前,人工智能系统与人类生产生活关系日益紧密,其安全问题越来越受到社会重视.国务院于2017年发布的《新一代人工智能发展规划》中明确指出:"在大力发展人工智能的同时,必须高度重视可能带来的安全风险挑战,加强前瞻预防与约束引导,最大限度降低风险,确保人工智能安全、可靠、可控发展"[25].遗憾的是,以往的人工智能理论大多基于一种"好人假设",较少考虑到在开放甚至是对抗环境下的机器学习安全与隐私问题.

从上述问题出发,本文结合当前人工智能系统安全领域的相关研究工作,系统地分析和归纳了人工智能系统中可能存在的安全与隐私风险及现有的应对方法,并对未来的发展趋势进行了展望,以期引起相关研究者的关注并提供指导.

1 人工智能系统安全风险模型

对于系统进行安全风险分析,首先需要建立安全风险模型.对此,本节首先对人工智能系统中潜在的攻击面进行简要分析,并从攻击能力和攻击目标2个角度建立攻击者模型.

1.1 人工智能系统攻击面

人工智能系统的应用场合和作用功能多样,例如无人驾驶、声音识别、机器翻译等,核心部分主要

包括数据和模型.如图 1 所示,根据数据流向,人工智能系统主要包含了 4 个关键环节[26]:

1) 输入环节.人工智能系统通过传感器(摄像头、麦克风、激光雷达、GPS 等)获取外部环境数据,或者通过直接读取文件获取数据.

2) 数据预处理环节.输入的原始数据需要经过格式转换、尺度变换、数据压缩等预处理工作,以满足机器学习模型输入格式要求,同时降低数据量以保证系统工作的实时性.

3) 机器学习模型.机器学习模型是人工智能系统的核心,即"大脑",主要包括训练和测试 2 个阶段.在训练阶段,机器学习模型利用预处理过的训练数据对模型参数进行调节,以提升对于特定任务的工作性能(通常用准确率、召回率等指标衡量).对于强化学习(reinforcement learning),还存在模型与环境的动态交互过程.当训练完成时,机器学习模型就进入了测试阶段.训练好的模型将根据输入提供相应的输出结果.

4) 输出环节.人工智能系统会以标签、置信度等多种形式给予输出,为后续的分类、决策等任务提供支持.

由于人工智能系统所处环境的开放性,输入、输出 2 个环节会直接暴露在攻击威胁环境中.在后续的介绍中将会看到,即使在预处理环节或机器学习模型被隐藏的情况下,攻击者仍然可以通过发送轮询样本的方式对系统内部结构进行推测并发动攻击.

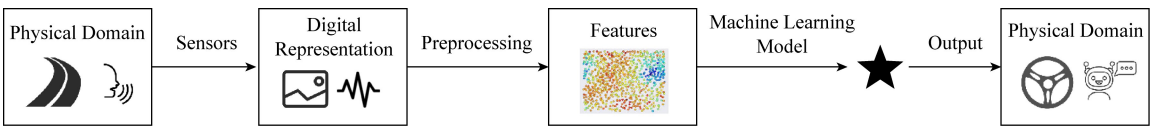

Fig. 1 The basic framework of artificial intelligence systems

图 1 人工智能系统基本框架[26]

1.2 攻击能力

在攻击者的攻击能力模型中,一般需要考虑 2 个要素:攻击者掌握的情报以及攻击者能够采取的攻击手段.

1) 依据攻击者掌握的情报,攻击可以分为:

① 白盒攻击(white-box attack).攻击者了解目标系统的详细信息,如数据预处理方法、模型结构、模型参数,某些情况下攻击者还能够掌握部分或全部的训练数据信息.在白盒攻击模型中,攻击者能够更容易地发现可攻击环节并设计相应的攻击策略.

② 黑盒攻击(black-box attack).系统对于攻击者而言并不透明,关键细节都被隐藏,攻击者仅能够接触输入和输出环节.在黑盒攻击模型中,攻击者可以通过构造并发送输入样本,并根据相应的输出信息来对系统的某些特性进行推理.

2) 依据攻击者能够采取的干扰手段,攻击被分为:

① 训练阶段攻击(attack in the training stage).攻击者可以干扰系统的训练阶段,主要方式包括对训练数据进行修改以及对环境施加影响(强化学习).

② 推断阶段攻击(attack in the inference stage).攻击者仅能接触到训练完成之后的系统.该假设在真实场景中更为多见.

1.3 攻击目标

攻击目标是指攻击者希望借助攻击所能达到的攻击效果.根据信息安全的 CIA 三要素,针对人工智能系统的攻击目标主要可以分为 3 类:

1) 保密性(confidentiality)攻击.攻击者期望从人工智能系统中盗取训练数据、模型参数等保密信息,破坏数据和模型隐私.

2) 完整性(integrity)攻击.攻击者期望能够影响系统输出,使其偏离预期.例如通过欺骗、篡改等攻击手段使得系统错误地接受假类样本,即错误接受(false acceptance).

3) 可用性(availability)攻击.攻击者期望降低系统的工作性能(如准确率)或者服务质量(如响应速度),甚至导致系统拒绝服务.

而根据攻击者的攻击效果,攻击目标又被划分为

1) 目标攻击(targeted attack).攻击者限定攻击范围和攻击效果,如诱导机器学习模型误分类到特定结果;

2) 无目标/无差别攻击(untargeted/indiscriminate attack).攻击者的攻击目标更为宽泛,攻击目的可能是造成更大的攻击影响,或者仅使得模型犯错而不限定欺骗结果.

本文将基于所建立的安全风险模型来审视人工智能系统在输入环节、数据预处理环节、机器学习模型以及输出环节 4 个核心模块,以及系统实现与运行中所面临的安全风险,并结合相关研究工作进行阐述和讨论.

2 输入环节安全风险及对策

人工智能系统依靠传感器(如摄像头、麦克风等)或数据文件输入(文件上传)获取信息,并通过数据预处理环节,依据模型输入要求将采集到的原始数据进行格式、大小等属性的调整.一旦攻击者借助某种方式对输入环节进行了干扰,就能够从源头上对系统发动攻击.传感器欺骗攻击即为一种典型的针对输入环节的缺陷利用.

1) 传感器欺骗.传感器欺骗是指攻击者针对传感器的工作特性,恶意构造相应的攻击样本并输送至传感器,造成人类和传感器对数据的感知差异,从而达成欺骗效果.该问题被认为是对配备有传感器的设备的最关键威胁之一,受到研究者的广泛关注.Shin 等人调查并将传感器欺骗攻击[27]分为 3 类:常规信道攻击(重放攻击)、传输信道攻击和侧信道攻击.传感器欺骗一个典型的例子是“无声”语音命令攻击.该类攻击借助人类听觉系统难以察觉的声音信号对语音识别系统开展攻击[28-29].对现代电子设备中普遍使用的非线性麦克风硬件而言,其可录制范围上限为 24 kHz,超越了人类对 20 kHz 可识别声音频率的上限.攻击者可以在麦克风超出人类听觉的接收频率范围内发送声音信号,从而使得设备能够感知而不被听众察觉.由于其不可闻性,该类攻击方法攻击效果更强.Zhang 等人提出的“海豚音攻击”通过生成超声频段的语音控制信号实现了对语音系统的“无声”控制[28].Dean 等人证明[30-31],当声频成分接近陀螺仪传感质量的共振频率时,MEMS 陀螺仪容易受到高功率高频声噪声的影响.攻击者可以借此干扰无人机等智能设备的环境感知能力,致使设备瘫痪.但是上述攻击假设攻击源可以在物理上靠近目标设备,难以实现远程攻击.

2) 应对措施.对于传感器欺骗攻击,可以采取传感器增强(忽略相应的攻击频段)、输入滤波等措施[28]来检测破坏恶意构造的攻击信息,实现对系统输入环节的安全增强.

3 数据预处理环节安全风险及对策

信息预处理环节是信息处理系统中的必备环节,其作用是将输入数据转换为后续模型输入所要求的特定形式.最近的研究表明,在数据的转换过程中也存在安全风险.

1) 重采样攻击.信息预处理环节的作用通常是为了将输入数据转换为模型输入要求的特定形式.数据重采样就是一种常见的数据预处理操作,其目的为:一是改变数据信息格式以满足输入要求,如当前主流视觉深度学习模型输入大小固定,需要对输入图片进行缩放操作;二是信息压缩,提升信息系统处理效率.这一过程会造成数据信息发生变化,成为一个潜在的攻击面.Xiao 等人提出了针对图像预处理环节的欺骗攻击[32],该方法是一种针对插值算法的逆向攻击方法,当攻击图片被图像识别系统缩放后,被隐藏图片得以显现.与经典的对抗样本攻击方法不同,该方法针对的是图像预处理环节,理论上与图像识别模型无关,并且该方法可以实现源-目标攻击(source-to-target attack).此外,该工作还显示,即使识别系统部署在云端,攻击者仍然可以通过轮询的方式对重采样过程进行推测和还原,进而发动重采样攻击.

2) 应对措施.针对重采样攻击,可以采取对输入预处理引入随机化或者重采样质量监测方法来增大攻击难度[32].

4 机器学习模型中的安全风险及对策

机器学习模型是人工智能系统进行感知和决策的核心部分,其应用过程主要包含训练和预测 2 个重要阶段.关于机器学习模型的安全问题,Dalvi 等人于 2004 年最早提出了对抗分类(adversarial classification)的概念[33],Lowd 等人于 2005 年进一步提出了对抗学习(adversarial learning)的概念[34].Huang 等人则对抗机器学习提出了更为具体和系统的分类方式[35].目前,机器学习模型安全问题可以主要分为 3 类:

1) 诱导攻击(causative attack).攻击者借助向训练数据加入毒化数据等手段,影响模型训练过程,进而干扰模型的工作效果.

2) 逃逸攻击(evasion attack).攻击者在正常样本基础上人为地构造异常输入样本,致使模型在分类或决策时出现错误,达到规避检测的攻击效果.

3) 探索攻击(exploratory attack).攻击者试图推断机器学习模型是如何工作的,包括对模型边界的预测、训练数据的推测等.

从保密性角度考虑,一般人工智能系统需要考虑 2 个要素——数据与模型.人工智能服务提供商

需要投入资金和时间收集数据,设计、训练和改进模型,同时需要对用户负责,保证数据不被泄露.然而,已有研究证明存在模型逆向攻击(model inversion attack)——可以根据系统输出推测输入特征,还原敏感信息[36-37],以及模型萃取攻击(model extraction attack)——通过发送轮询数据推测模型参数并尝试还原出功能相近的替身模型(substitution model)[38],二者会分别侵犯数据隐私和模型隐私.

4.1 数据投毒

数据投毒是指攻击者通过修改训练数据内容和分布,来影响模型的训练结果.例如 Yang 等人展示了攻击者通过对推荐系统注入构造的虚假关联数据,污染训练数据集,实现对推荐系统反馈结果的人为干预[39].实验表明:通过对共同访问(co-visitation)推荐系统进行数据投毒,可以对 YouTube,eBay,Amazon,Yelp,LinkedIn 等 Web 推荐系统功能产生干扰.Munñoz-González 等人提出了基于反向梯度优化的攻击方法,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,都可以实现数据投毒效果[40].

应对措施.针对投毒攻击的防御,一般考虑污染数据和正常数据分布差异,方法主要包括鲁棒性机器学习[41]以及数据清洗[42].

4.2 模型后门

模型后门(backdoor)是指通过训练得到的、深度神经网络中的隐藏模式.当且仅当输入为触发样本(trigger)时,模型才会产生特定的隐藏行为;否则,模型工作表现保持正常.Gu 等人提出了 BadNets,通过数据投毒方式来注入后门数据集[43].针对 MNIST 手写数据集识别模型网络,使用 BadNets 可以达到 99%以上的攻击成功率,但不会影响模型在正常手写样本上的识别性能.Liu 等人提出了一种针对神经网络的特洛伊木马攻击[44].相较于 Gu 等人的工作[43],该方法的一个优点是攻击者无需直接接触训练集.该方法另一个优点在于在触发样本和神经元之间构建了更强的连接,在训练样本较少的情况下也能够注入有效后门.然而,该后门构造方法基于引起特定最大响应值的内部神经元来设计触发样本,无法构造任意触发样本.除了触发模型的异常行为外,Song 等人展示了一种利用模型后门的训练数据隐私窃取攻击方式[45]:依靠类似机器学习中正则化或数据增强方法对机器学习算法进行微调,第三方机器学习服务提供商可以借助用户数据训练出高准确度和高泛化性能的模型,并使得该模型能够暴露训练数据信息.

针对模型后门问题,Wang 等人提出了相应的检测和后门还原方案[46],该方法的思想相对直观:对于模型后门相对应的标签,很小的输入扰动会引起该标签对应置信度明显的变化.此外,作者还提出了包括输入过滤、神经元裁剪以及去学习等后门去除策略.

4.3 对抗样本

传统机器学习模型大多基于一个稳定性假设:训练数据与测试数据近似服从相同分布.当罕见样本甚至是恶意构造的非正常样本输入到机器学习模型时,就有可能导致机器学习模型输出异常结果.一个典型例子即 Szegedy 等人在 2013 年所描述的视觉"对抗样本"(adversarial examples)现象: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的结果[47].通过构造对抗样本,攻击者可以通过干扰人工智能服务推理过程来达成逃避检测等攻击效果.

在机器视觉领域,针对对抗样本的生成方法和对抗样本特性已得到较多的研究.根据攻击效果分类,对抗样本攻击可以被分类为目标攻击[48]和非目标攻击[49],而根据攻击者对机器学习模型的攻击能力则可以将攻击分类为白盒攻击[50](white-box attack)和黑盒攻击[51](black-box attack).为了达到欺骗效果,对抗样本的一个显著特点是隐蔽性,即对抗扰动难以被人类所察觉,最大限度保持原样本的语义信息.除了隐蔽性之外,Tramèr 等人的工作[52]还揭示了对抗样本的另一个突出特性——可传递性(transferability).借助可传递性,同一个对抗样本可以同时作用于多个模型,这部分解释了对抗样本问题为什么得以广泛存在.可传递性使得对抗样本防御工作变得更具挑战性.针对目前常见的对抗样本生成算法,一些研究人员合作发布了开源对抗样本算法库 CleverHans[53],以推动对抗样本攻/防研究工作的发展.与此同时,一些研究者探究了物理世界中的对抗样本现象:Kurakin 等人尝试了将对抗样本进行打印[54].针对 Inception-V3 模型生成对抗样本并利用 600 像素分辨率的打印机进行打印,打印出的对抗样本对于识别系统仍然具有欺骗性.然而该方法需要在所打印出的图片四周配备二维码以帮助图像识别系统对图像进行定位和裁剪,其结果并不具有普适性;Athalye 等人提出了鲁棒性更高的

物理对抗样本生成技术,利用 3D 打印技术制作了物理世界对抗样本模型,可以在多个角度下实现对识别模型的欺骗[55];Sharif 等人引入不可打印分数(non-printability score, NPS)到目标函数中,通过优化方法计算扰动并打印到眼镜框上,攻击者佩戴眼镜框后可以成功误导人脸识别系统[56];Eykholt 等人则综合考虑了相机角度、干扰形状、打印效果等物理因素,设计了针对无人驾驶系统的对抗样本攻击方法,通过对路标覆盖扰动标记,诱导无人驾驶系统将"停车"标志被误识别为"限速"标志[57].这些例子进一步说明了对抗样本威胁不仅局限于信息域,对抗样本攻击能够在物理域产生实际影响,在一些关键应用上可能会引发灾难性后果.

在语音系统方面,Kumar 等人开展了针对语音错误解释攻击的实证研究[58].此外,Zhang 等人展示了一种类似的方法[59]:针对语音助手的"技能"(skill,即某种功能)调用方式,攻击者通过引入具有相似/部分覆盖的发音名称或释义名称的恶意"技能",来劫持目标"技能"的语音命令.Zhang 等人则系统地探究了自然语言处理和意图分类器(intent classifier)的工作归因,并创建了第 1 个语言模型引导的模糊测试工具,以发现现有明显更易受攻击的语音应用[60].除了利用自然语言理解缺陷的语音系统攻击,一些研究人员提出了一系列的语音对抗样本生成方法,来欺骗语音识别系统.Carlini 等人开发了一种针对 Mozilla DeepSpeech 的对抗音频生成技术,利用优化方法直接对原始输入进行修改从而对模型进行欺骗[61].Yuan 等人提出针对服务接口的对抗语音样本生成方法,攻击者将一组命令嵌入到一首歌中,可以有效地控制目标系统而不被察觉[62].Vaidya 等人提出的方法利用合成和自然声音之间的差异,制造可以被计算机语音识别系统识别但人类不易理解的对抗样本[63].Carlini 等人的工作展示了利用一种迭代方法来构造针对黑盒语音系统的攻击语音[64].为了获得更好的结果,Carlini 等人同时提出了一种针对白盒模型的改进攻击方法.在攻击者完全了解语音识别系统中所使用算法的条件下,这种改进后的攻击可以保证合成的语音命令不被人类所理解.在威胁模型假设方面,该类攻击要求将攻击者的发言者放置在受害者设备附近的物理位置(距离超过 3.5 m 时会失效).上述 4 个攻击方法局限于特定的模型和硬件平台.不同于此,Abdullah 等人提出了一种针对声音处理环节的攻击方法[65].该文作者提出了 4 种扰动类型,并在包括 Google 语音 API,Bing 语音 API 等7 种语音服务在内的 12 个语音识别模型上进行了测试,均成功实现了有效攻击,展示了该攻击影响的广泛性.

除了视觉系统与语音系统外,文本处理系统也是人工智能技术的一类典型应用,被广泛应用于垃圾邮件检测、不良信息过滤、机器翻译等任务上.当前研究表明文本处理系统也正受到对抗样本的威胁.Papernot 等人提出了一种基于梯度的白盒对抗样本生成方式[66],该方法迭代地修改输入文本,直到生成的序列被循环神经网络错误分类,但该攻击引发的词级变化会明显影响文本语义,攻击容易被察觉;Samanta 等人利用嵌入梯度来确定重要单词[67],并设计了启发式规则、人工构造的同义词及笔误来对文本进行删除、增加或替换;Ebrahimi 等人提出了一种基于梯度的字符级分类器对抗样本构造方法,对 one-hot 编码形式的输入向量进行修改[68];Alzantot 等人提出了同义词替换攻击方法[69],利用遗传算法生成使用同义词或近义词替换的方法,通过对抗性文本来欺骗语义识别系统;Belinkov 等人的研究表明字符级的机器翻译系统对数据噪声十分敏感,可以借助非词汇符号进行攻击[70];同样地,Gao 等人提出一种黑盒文字对抗样本攻击方法,应用字符扰动来生成针对深度学习分类器的对抗性文本[71];Hosseini 等人的工作展示,通过在字符之间添加空格或点号,就可以彻底改变 Google 有害信息检测服务的评分[72];Zhao 等人还提出了利用生成对抗网络生成针对机器翻译应用程序的对抗序列[73],然而该方法仅限于短文本.

除了上述研究工作外,还存在针对其他应用类型的对抗样本攻击.Xu 等人利用遗传编程(genetic programming)方法随机修改文件,成功攻击了 2 个号称准确率极高的恶意 PDF 文件分类器:PDFrate 和 Hidost[74].这些逃避检测的恶意文件都由算法自动修改生成,并不需要 PDF 安全专家介入.在恶意代码检测方面,Grosse 等人提出了在离散和二进制输入域修改输入样本,可以绕过恶意有效代码检测[75].

为了应对对抗样本的问题,近年来研究人员提出了一些包括直接对抗训练[47](adversarial training)——将对抗样本及正确标签重新输入到模型中进行重训练,该方法较为简单但防御未知对抗样本能力较差;梯度掩模[76](gradient masking)——针对基于梯度的对抗样本攻击方式,通过隐藏梯度,

令此类攻击失效;对抗样本检测[77]——直接检测是否存在对抗样本的防御方法.此外,Dziugaite 等人使用 JPG 图像压缩的方法,减少对抗扰动对准确率的影响[78].实验证明该方法对部分对抗攻击算法有效,但通常仅采用压缩方法是远远不够的,并且压缩图像时也会降低正常分类的准确率.

虽然对抗样本防御方法已经得到较多研究,但是当前仍然缺少一个通用有效的防御方案.事实上,当前大多数的防御评估方法都是在衡量对抗攻击的能力下界[79]:这类评估所验证的是一个样本集合的邻域内的攻击样本攻击效果,仅能发现当前区域而非所有防御失效点.而且这些防御评估方法都是基于一种非适应性攻击模型,即假设攻击者并不知晓防御方法.Carlini 等人认为考虑非适应性攻击模型是有必要的,但是有很大的局限性[80].相对应地,一种有效的模型鲁棒性评估应该基于适应性攻击模型,即假设攻击者知晓防御者已采取的防御策略并可以采取反制措施[81-82].例如针对梯度掩模防御策略,Papernot 等人提出了一种通过黑盒轮询输入标签的策略来对梯度进行回推[51];Athalye 等人则提出了通过改变代价函数来进行对抗样本攻击[83].从安全评估结果的可靠性考虑,需要对所有已知或未知攻击(最坏情况)的防御效果得出被测试模型的鲁棒性下界,即模型鲁棒性的最低保证.

当前一个发展方向是对模型鲁棒性进行形式化验证.虽然 Lecuyer 等人的研究成果可以应用于对 ImageNet 分类器的鲁棒性分析[84],但当前的模型鲁棒性验证方法,如文献[85-86],还大多只能局限于特定的网络模型.文献[87-88]等工作已经开始探索对任意神经网络模型鲁棒性进行形式化验证的可能性,但是由于计算复杂度过高,无法应用于中大规模的网络模型.此外,鲁棒性证明方法的一个显著缺点是,该类方法给出了对于特定集合的邻域对抗样本存在性证明,但是尚无法对该集合外的样本提供理论上的证明和保证[80].

4.4 模型逆向

由于机器学习模型在训练时会或多或少地在训练数据上发生过拟合,攻击者可以根据训练数据与非训练数据的拟合差异来窥探训练数据隐私.Fredrikson 等人以医疗机器学习中的隐私问题为例阐述了模型逆向攻击(model inversion attack)[89]:对某一个被训练好的机器学习模型,攻击者利用模型、未知属性以及模型输出的相关相关性,实现对隐私属性的推测.具体到实例中,Fredrikson 等人根据华法林剂量信息来尝试对患者的基因型进行推测.此外,Fredrikson 等人在[36]展示了针对另外 2 个模型进行逆向攻击的例子:借助模型置信度输出,攻击者可以估计生活调查中的受访者是否承认对其他重要人物存在欺骗行为;针对人脸识别系统,攻击者可以根据用户姓名恢复出对应的可识别的人脸照片.一些研究证明了另外一类的模型逆向攻击——成员推理攻击(membership inference attack),即攻击者可以推断某个特定实例是否在训练数据集中.早在 2008 年 Homer 等人就展示了对基因组数据的成员推理攻击(membership attack)[90].在此基础上,Shokri 等人展示了可以通过训练多个“影子模型”(shadow models)来模拟被攻击模型,并利用机器学习模型输出中暗含的训练数据之间的区分性,来发动成员推理攻击[37].Salem 等人通过实验证明了通过单个影子模型开展相同攻击的可能性[91],即使在攻击者无法获取被攻击模型的训练数据情况下,也根据模型输出的统计特征进行推测攻击.针对地理位置聚集信息,Pyrgelis 等人建立了博弈模型,并将其转化为是否属于特定集合成员的分类问题,进一步实现了对于地理位置信息的成员推理攻击[92].除了判别模型,Hayes 等人还提出了白盒和黑盒情况下针对于生成模型的成员推理攻击,并在多个数据集上开展了实证研究[93];Salem 等人提出了针对在线学习算法的数据重构攻击,对在线学习模型的更新训练数据进行了推测和复原[94].以往大多数相关研究工作会采用一些攻击者拥有同分布数据、影子模型或者目标模型结构等假设.对此 Salem 等人研究了这些假设逐步弱化时的成员推理攻击情况[91].结果表明,即使在已知信息很有限的情况下,攻击者仍然具有进行成员推理攻击的能力.

除此之外,Carlini 等人揭露了深度学习模型,尤其是生成模型中存在的“意外记忆问题”[95]——模型在对低频的敏感训练数据(如用户密码等)进行学习的同时,会倾向于完整地记忆与目标任务无关的训练数据细节,这就为该类数据带来了泄露风险.实验结果表明,传统的过拟合抑制方法很难解决意外记忆问题.对此,Carlini 等人提出了对应的“暴露度”(exposure)指标来评估意外记忆程度,用于辅助开发者进行模型结构和参数的选择、调整.

针对用户数据保护问题,研究者提出了多种解决方案.常见的一种方法是利用差分隐私(differential

privacy)模型[96]来分析算法所能提供的隐私性保证.Chaudhuri 等人证明在训练时通过向代价函数,即模型预测值与标签的误差加入指数分布的噪声,可以实现 ε-差分隐私[97].Abadi 等人提出在梯度被用于参数更新前对梯度添加扰动,可以达到单一训练方场景下的一种强差分隐私边界[98].Shokri 等人证明对于类似深度神经网络的大容量模型,借助引入噪声参数的多方计算,可以保证差分隐私性[99].Gilad-Bachrach 等人提出了一种神经网络模型的加密方法——CryptoNets,该方法使得神经网络可以被应用于加密数据[100].CryptoNets 允许用户向云端服务上传加密数据,而无需提供秘钥,从而保证了用户数据的机密性.为了保证用户数据隐私,拥有训练数据的双方或者多方可能不被允许直接进行训练数据的交换和合并,这就造成了"数据孤岛"问题.对此,有研究提出利用联邦学习(federated learning)方法来进行多方联合学习[101].在该模型下,训练数据并不会离开本地.各方建立一个虚拟共有模型,通过加噪机制交换参数,对共有模型进行共同训练.

4.5 模型萃取

模型萃取攻击(model extraction attack)是指攻击者可以通过发送轮询数据并查看对应的响应结果,推测机器学习模型的参数或功能,复制一个功能相似甚至完全相同的机器学习模型.例如理论上讲,针对 n 维线性回归模型,通过 n 组线性不相关轮询数据及模型输出可准确求解出权重参数[38].该攻击可破坏算法机密性,造成对知识产权的侵犯,并使攻击者随后能够依据被复制模型进行对抗样本攻击或模型逆向攻击.Lowd 和 Meek 提出了有效的算法来窃取线性分类器的模型参数[34].Tramèr 等人证明,当 API 为返回置信度分数时,可以更准确和有效地推测模型参数[38].此外,超参数在机器学习中至关重要,因为超参数的差异通常会导致模型具有显著不同的性能.根据机器学习模型最终学习到的参数往往会最小化代价函数这一原则,Wang 等人提出了机器学习模型的超参数推测方法[102].

对于模型萃取攻击,最直接最简单的防御策略是对模型参数[102]或者输出结果进行近似处理[38].除此之外,为了避免模型被盗用、保护知识产权,一些研究者还提出了模型水印(watermarking)的概念.Venugopal 等人较早地提出关于学习模型水印技术的方法[103],但是它侧重于标记模型的输出而非标记模型本身.文献[104-105]提出通过向损失函数添加新的正则化项来对神经网络添加水印的方法.虽然他们的方法保持了模型的高识别精度,同时使水印具有一定的抗毁能力,但其并没有明确解决所有权的虚假声明问题,也没有明确考虑水印生成算法遭泄露后的抗攻击情况.此外,在文献[104-105]中,为了避免因密钥泄露而发生的水印移除情况,验证密钥只能使用一次,这带来了一定的局限性.Merrer 等人建议结合对抗样本与对抗训练方法为神经网络注入水印[106].他们提出生成 2 种类型(被模型正确和错误地分类)的对抗样本,然后微调模型以使其正确地对所有类型进行分类.这种方法在很大程度上依赖于对抗样本以及它们在不同模型中的可迁移性,但目前尚不明确对抗样本在什么条件下能够进行跨模型迁移,或者这种迁移性是否会被削弱[107].Adi 等人提出了一种黑盒方式的深度神经网络水印技术[108],从理论上分析了该方法与模型后门的联系,并通过实验证明了该方法不影响原模型性能,同时对水印的鲁棒性进行了评估.

5 输出环节安全风险

模型输出将会直接决定人工智能系统的分类和决策.通过对决策输出部分的劫持和结果篡改可以直接实现对系统的干扰或控制.另一个需要注意的问题是,多数人工智能服务接口会反馈丰富的信息,但是丰富/准确的决策输出值可能会带来安全隐患——攻击者据此可以开展模型逆向攻击和模型萃取攻击,或者利用置信度来迭代式构造对抗样本.此外,Elsayed 等人介绍了一种对抗性重编程方法(adversarial reprogramming)[109].即使模型的训练目的并非是完成攻击者所指定的任务,攻击者通过制造一个对抗扰动并添加至机器学习模型的所有测试输入,可以使模型在处理这些输入时执行攻击者选择的任务.利用该方法,攻击者只需要付出很小的代价,就可以借助他人训练好的模型资源实现所需的系统功能.

如 4.4 节和 4.5 节所述,针对利用模型输出置信度进行数据逆向、模型萃取或模型重用等探索攻击行为,可以采用输出值近似处理或引入随机波动来降低探索攻击反馈结果的准确性,提高攻击难度.

6 系统实际搭建及运行中的安全风险及对策

6.1 代码漏洞

当前流行的深度学习框架,如 Caffe,Tensorflow,Torch 等,提供了高效、便捷的人工智能系统开发支持环境,为人工智能技术的推广作出了巨大贡献.仅需几百行甚至几十行的核心代码就可以完成模型的搭建、训练和运行.但与框架的使用简洁性恰恰相反,为了完成对多种软硬件平台的支持以及复杂计算功能的集成,深度学习框架往往需要依赖于种类纷繁的基础库和第三方组件支持,如 Caffe 包含有超过 130 种的依赖库[110].这种组件的依赖复杂度会严重降低深度学习框架的安全性.某个组件开发者的疏忽,或者不同组件开发者之间开发规范的不统一,都可能会向深度学习框架引入漏洞.更为严重的是,一个底层依赖库的漏洞(如图像处理库 OpenCV)有可能会蔓延到多个高层深度学习框架,进而影响到所支持的一系列应用中.此时攻击者可以基于控制流改写人工智能系统关键数据,或者通过数据流劫持控制代码执行,实现对人工智能系统的干扰、控制甚至破坏.Xiao 等人分析了深度学习应用的层级结构,并披露了 Tensorflow,Caffe 与 Torch 三种深度学习框架及其依赖库中的数十种代码漏洞,同时展示了如何利用该漏洞引发基于 3 种框架的深度学习应用发生崩溃、识别结果篡改、非法提权等问题[110].

通常来说,可以利用传统的漏洞测试方法,例如模糊测试来发现软件中的代码漏洞.但是,传统的漏洞测试方法在应用于深度学习框架时具有其局限性.Xiao 等人指出,基于覆盖率的模糊测试方法对于深度学习应用的测试效果并不理想[110].其原因在于基本上所有的输入数据都经过相同的网络层进行计算,导致大量输入样本覆盖的是同一条执行路径.另一个问题则在于难以区分代码逻辑漏洞和模型本身的对抗样本/训练不完全问题.

6.2 学习不完全/学习偏差

虽然依靠海量数据和深度学习技术的人工智能系统在多种任务上表现出突出的工作性能,但由于诸如训练数据偏差、过拟合和模型缺陷等原因,即便在非对抗环境下,系统罕见或边缘样本输入可能会引起人工智能系统出现意外或错误的行为.例如自动驾驶训练数据无法覆盖所有光照、天气、道路及周围物体分布下的行驶情况,致使未知路况下无人驾驶汽车行为的准确性和可预测性难以得到保证.在安全要求较高的场合,罕见/边缘样本的训练缺失有可能导致灾难性后果.由于目前深度学习模型可解释性差,难以对系统异常行为进行预测或归因,发现由训练不完全或偏差导致的模型缺陷成为一个极具挑战性的问题.

为了发现人工智能系统中潜藏的漏洞,相关研究工作将软件自动化测试中的概念迁移到了人工智能领域.Pei 等人设计了深度学习系统的白盒测试框架 DeepXplore[111],并提出了“神经元覆盖率(neuron coverage)”的概念,该框架会按照一定策略自动生成测试样本来触发潜在的异常行为,以帮助发现网络缺陷;Ma 等人在文献[111]基础上提出了多方位细粒度的自动化测试方法 DeepGauge[112],并提出了更详细的神经网络自动化测试指标;受 MC/DC 测试覆盖率指标的启发,Sun 等人提出了基于 DNN 结构特征和语义的 4 种测试指标[113],并在 MNIST,CIFAR-10 和 ImageNet 分类任务上进行了验证测试;Ma 等人将传统软件测试中的组合测试(combinatorial testing)概念延伸到深度学习模型上并提出了 DeepCT[114];Ma 等人还将变异测试(mutation testing)概念沿用到深度学习模型上并提出了 DeepMutation 测试框架[115],设计了针对训练数据和训练过程的原始级变异方法,以及针对无训练环节的模型级变异方法;针对神经网络在数值传递过程中可能存在的漏洞,Odena 等人提出了针对神经网络的基于覆盖指导的模糊(coverage-guided fuzzing)方法 TensorFuzz,以帮助代码调试[116].除了自动化测试之外,还有学者尝试了形式化分析方法:Wang 等人基于区间型符号的神经网络形式化安全分析方法[117],根据输入估计网络的输出范围,判断是否会触犯某些安全限定.

6.3 系统设计缺陷利用

为了提高系统的智能化,诸如语音助手等的人工智能服务需要被赋予很高的系统操作权限,一旦设计不当,很容易被攻击者利用进行系统非法操作.例如 Diao 等人展示了攻击者可以控制设备扬声器,在后台播放准备好的音频文件,同时借助安卓系统内嵌的谷歌语音助手,进行无权限情况下的发送信息、读取隐私数据、甚至是远程控制等操作[118].

综上对人工智能系统中的安全风险进行总结,如表 1 所示:

Table 1 A Brief Summary of Security and Privacy Risks Against Artificial Intelligence Systems

表 1 人工智能系统安全与隐私风险分析小结

Pipeline	Type	Threat Model	Description	Instances	Impacts
Data Input	Sensor Spoofing	Inference/White-box or Black-box/Impact integrity or availability	Leverage the sensibility difference between human and hardware to inject deceiving/malicious information into inputs which are hard to be perceived by human.	Ref [28-29]	Spoofing
Data Preprocessing	Scaling Attack	Inference/White-box or Black-box/Impact integrity	Conceal a tiny amount of deceiving/malicious information under a large amount of normal information, and the injected part will get recovered after resampling.	Ref [32]	Spoofing, detection evasion
Machine Learning Model	Data Poisoning	Training/White-box/Impact integrity	Tampering the content or distribution of the training data.	Ref [39-40]	Destroyor control model functionality
	Backdoor	Training/White-box/Impact integrity	A hidden pattern trained into a model, which can be activated and produce unexpected behavior if and only if when specific "triggers" get input.	Ref [43,45]	Hide unexpected behavior
	Adversarial Examples	Inference/White-box or Black-box/Impact integrity	Applying small but intentionally designed perturbations to test samples, which can cause the models to give incorrect outputs.	Ref [47,70]	Spoofing, detection evasion
	Model Inversion	Inference/Black-box/Impact(data) confidentiality	For an already-trained model, one adversary can infer private or sensitive attributes by leveraging the correlation among model, hidden attributes and model outputs.	Ref [36-37, 89,92]	Training data leakage
	Unintended Memory	Inference/White-box/Impact (data) confidentiality	When trained on rarely-occurring data, the model tends to remember too much details even unrelated to the learning task.	Ref [95]	Sensitive data leakage
	Model Extraction	Inference/Black-box/Impact (model) confidentiality	Steal parameters of a machine learning model by sending queries, to construct a new model with similar functionality and performance.	Ref [34,38, 102]	Partial or complete model functionality duplication, which enables black-box attacks
Output	Model Reuse	Inference/Black-box/impact (model) confidentiality	Add perturbation to all testing-time inputs to reprogram neural networks to perform a specific task chosen by the adversary.	Ref [109]	Steal and transfer model functionality
Implementation Or Execution	Code Vulnerability	Inference/White-box/Impact integrity or availability	Vulnerabilities buried in codes of machine learning libraries or systems, which can be propagated and spread following the "third-party dependencies → deep learning frameworks → AI systems" path.	Ref [110]	Spoofing, hijacking, denial of service and etc.
	Incomplete/Biased Learning	Inherent weaknesses, which may affect integrity or availability	Unexpected or incorrect output due to underfitting, overfitting or biased training data.		Unexpected or incorrect behavior
	System Design Flaw	Inference/Black-box/Impact integrity or availability	Inappropriate system design in logic or permission control.	Ref [118]	Remote sensitive data access or control without any permission

7 人工智能安全分析与防护技术的研究展望

针对表 1 中所总结的人工智能系统安全与隐私问题,在本节,我们将讨论在人工智能安全分析与防护研究工作中的 4 个发展方向:

1) ~~物理对抗样本.针~~对无人驾驶、人脸识别、语音识别等关键应用,需要评估其在真实场景下的安全性能,尤其是潜在的物理对抗样本威胁.不同于信息域内对图像、音频等文件直接进行修改的对抗样本攻击方式,物理对抗样本攻击效能评估还需要同时考虑物理环境以及输入输出设备特性等因素的影

响.例如针对视觉系统而言,还需考虑光照、角度、摄像头光学特性、打印设备分辨率及色差等因素对构造对抗样本的影响;对音频处理系统而言,进行音频对抗样本的重放攻击需同时考虑攻击扬声器的声音播放质量、目标麦克风的收音性能以及背景噪声等因素的影响.

2) 模型鲁棒性的形式化验证.形式化验证可以给出对于攻击的上界/模型鲁棒性下界的估计,对于安全系数要求较高的场合而言是十分必要的.可以预见,形式化验证将是今后模型安全评估的一个重要研究方向,会有越来越多的研究集中在如何降低验证复杂度以及提高方法的模型普适性上.

3) 人工智能系统自动化测试方法.当前形式化验证方法计算复杂度高、难以应用到实际深度模型上.此外,复杂的代码依赖层级给人工智能系统的人工分析带来极大的难度.对此,可以借助自动化测试方法来持续提高对攻击强度的平均估计,发现模型可能出现的异常行为或者安全漏洞.除了代码自动化测试方法以外,模型的自动化测试也可以作为模型形式化验证的一种辅助措施.在设计和应用自动化测试方法时需要关注 3 个问题:①如何定义模型异常行为;②如何区分模型在无意义分类边界下和关键分类边界下的异常行为;③如何定义自动化评测的引导指标.

4) 隐私保护.在某些应用场景中,相较于人工智能服务的精度,用户更重视个人数据的隐私保护.尤其在大规模分布式数据存储和模型训练的情况下,如何同时保证用户数据隐私和模型的训练效率及工作精度是在人工智能服务提供商需要解决的关键问题.

8 结 论

随着深度学习技术及计算硬件架构的发展和变革,人工智能技术在机器视觉、语音识别、机器视觉等关键任务上取得了重大突破,接近甚至超过人类水平,这些成果推动了人工智能技术的技术落地,衍生出诸如人脸识别、语音助手、无人驾驶等应用领域.在促进人工智能系统为人类生产生活带来便利的同时,如何发现、修复人工系统中的安全缺陷,规避人工智能应用风险也成为了人类和社会日渐关心的问题.本文在对国内外人工智能安全研究调研和分析的基础上,总结归纳了数据输入、数据预处理、学习模型与模型输出 4 个系统关键点中可能存在的安全风险及应对措施,并进一步指出了人工智能安全分析与防护技术未来的研究趋势.

参 考 文 献

[1] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[2] Xiong W, Droppo J, Huang Xuedong, et al. Achieving human parity in conversational speech recognition [J]. arXiv preprint arXiv:1610.05256, 2016

[3] Fan Zhengguang, Qu Dan, Yan Honggang, et al. Fast incremental outlier mining algorithm based on grid and capacity [J]. Journal of Computer Research and Development, 2017, 54(5): 1036-1044 (in Chinese)
(范正光, 屈丹, 闫红刚, 等. 基于深层神经网络的多特征关联声学建模方法[J]. 计算机研究与发展, 2017, 54(5): 1036-1044)

[4] Yu A W, Dohan D, Luong M T, et al. QANet: Combining local convolution with global self-attention for reading comprehension [J]. arXiv preprint arXiv:1804.09541, 2018

[5] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550: 354-359

[6] Stoica I, Song D, Popa R A, et al. A Berkeley view of systems challenges for AI [J]. arXiv preprint arXiv: 1712.05855, 2017

[7] Jia Yangqing, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding [C] // Proc of the 22nd ACM Int Conf on Multimedia. New York: ACM, 2014: 675-678

[8] Abadi M, Barham P, Chen Jianmin, et al. Tensorflow: A system for large-scale machine learning [C] //Proc of the 12th USENIX Symp on Operating Systems Design and Implementation (OSDI'16). Berkeley, CA: USENIX Association, 2016: 265-283

[9] Ronan, Clément, Koray, et al. Torch7 [EB/OL]. [2019-05-25]. http://torch.ch/

[10] Chen Tianqi, Li Mu, Li Yutian, et al. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems [J]. arXiv preprint arXiv: 1512.01274, 2015

[11] PaddlePaddle developers. PaddlePaddle [EB/OL]. [2019-05-25]. https://github.com/paddlepaddle/paddle

[12] Tencent. Tencent AI open platform [EB/OL]. [2019-05-25]. https://ai.qq.com/

[13] Alibaba. ET brain [EB/OL]. [2019-05-25]. https://et.aliyun.com/index

[14] Baidu. Baidu AI open platform [EB/OL]. [2019-05-25]. http://ai.baidu.com/

[15] Google. AI and machine learning products [EB/OL]. [2019-05-25]. https://cloud.google.com/products/ai/

[16] Azure. Azure AI [EB/OL]. [2019-05-25]. https://azure.microsoft.com/en-us/overview/ai-platform/

[17] Amazon. Amazon machine learning [EB/OL]. [2019-05-25]. https://aws.amazon.com/cn/machine-learning/

[18] IBM. IBM watson [EB/OL]. [2019-05-25]. https://www.ibm.com/watson

[19] Wakabayashi D. Self-driving uber car kills pedestrian in Arizona, where robots roam [EB/OL]. [2019-04-28]. https://www. nytimes. com/2018/03/19/technology/uber-driverless-fatality.html

[20] Wikipedia. Tay (bot) [EB/OL]. [2019-04-29]. https://en.wikipedia.org/wiki/Tay_(bot)

[21] Bojarski M, Testa D D, Dworakowski D, et al. End to end learning for self-driving cars [J]. arXiv preprint arXiv:1604.07316, 2016

[22] Wang Juanjuan, Qiao Ying, Wang Hongan. Graph-based auto-driving reasoning task scheduling [J]. Journal of Computer Research and Development, 2017, 54(8): 1693-1702 (in Chinese)
(王娟娟, 乔颖, 王宏安. 基于图模型的自动驾驶推理任务调度[J]. 计算机研究与发展, 2017, 54(8): 1693-1702)

[23] Yuan Zhenlong, Lu Yongqiang, Wang Zhaoguo, et al. Droid-Sec: Deep learning in android malware detection [C] //Proc of the 2014 ACM Conf on SIGCOMM. New York: ACM, 2014: 371-372

[24] Metz R. Using deep learning to make video surveillance smarter [EB/OL]. 2015 [2019-05-05]. https://www.technologyreview.com/s/540396/using-deep-learning-to-make-video-surveillance-smarter/

[25] The State Council. New-generation artificial intelligence development plan [EB/OL]. 2017 [2019-05-01]. http://www. gov. cn/zhengce/content/2017-07/20/content _ 5211996. htm (in Chinese)
(国务院. 新一代人工智能发展规划 [EB/OL]. 2017 [2019-05-01]. http://www. gov. cn/zhengce/content/2017-07/20/content_5211996.htm)

[26] Papernot N, McDaniel P, Sinha A, et al. Towards the science of security and privacy in machine learning [J]. arXiv preprint arXiv:1611.03814, 2016

[27] Shin H, Son Y, Park Y, et al. Sampling race: Bypassing timing-based analog active sensor spoofing detection on analog-digital systems [C] //Proc of the 10th USENIX Workshop on Offensive Technologies (WOOT'16). Berkeley, CA: USENIX Association, 2016

[28] Zhang Guoming, Chen Yan, Ji Xiaoyu, et al. DolphinAttack: Inaudible voice commands [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security (CCS'17). New York: ACM, 2017: 103-117

[29] Roy N, Hassanieh H, Roy Choudhury R. BackDoor: Making microphones hear inaudible sounds [C] //Proc of the 15th Annual Int Conf on Mobile Systems, Applications, and Services. New York: ACM, 2017: 2-14

[30] Dean R N, Castro S T, Flowers G T, et al. A characterization of the performance of a MEMS gyroscope in acoustically harsh environments [J]. IEEE Transactions on Industrial Electronics, 2010, 58(7): 2591-2596

[31] Dean R N, Flowers G T, Hodel A S, et al. On the degradation of MEMS gyroscope performance in the presence of high power acoustic noise [C] //Proc of the 2007 IEEE Int Symp on Industrial Electronics. Piscataway, NJ: IEEE, 2007: 1435-1440

[32] Xiao Qixue, Chen Yufei, Shen Chao, et al. Seeing is not believing: Camouflage attacks on image scaling algorithms [C] //Proc of the 28th USENIX Security Symp (USENIX Security'19). Berkeley, CA: USENIX Association, 2019

[33] Dalvi N, Domingos P, Sanghai S, et al. Adversarial classification [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 99-108

[34] Lowd D, Meek C. Adversarial learning [C] //Proc of the 11th ACM SIGKDD Int Conf on Knowledge Discovery in Data Mining. New York: ACM, 2005: 641-647

[35] Huang L, Joseph A D, Nelson B, et al. Adversarial machine learning [C] //Proc of the 4th ACM Workshop on Security and Artificial Intelligence. New York: ACM, 2011: 43-58

[36] 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 (CCS'15). New York: ACM, 2015: 1322-1333

[37] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models [C] //Proc of the 2017 IEEE Symp on Security and Privacy (S&P'17). Piscataway, NJ: IEEE, 2017: 3-18

[38] 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'16). Berkeley, CA: USENIX Association, 2016: 601-618

[39] Yang Guolei, Gong N Zhenqiang, Cai Ying. Fake co-visitation injection attacks to recommender systems [C] //Proc of the 24th Annual Network and Distributed System Security Symp (NDSS 2017). Reston, VA, USA: The Internet Society, 2017

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

[41] Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning [C] //Proc of the 2018 IEEE Symp on Security and Privacy (S&P'18). Piscataway, NJ: IEEE, 2018: 19-35

[42] Cretu G F, Stavrou A, Locasto M E, et al. Casting out demons: Sanitizing training data for anomaly sensors [C] // Proc of the 2008 IEEE Symp on Security and Privacy (S&P'08). Piscataway, NJ: IEEE, 2008: 81-95

[43] Gu Tianyu, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain [J]. arXiv preprint arXiv:1708.06733, 2017

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

[45] 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 (CCS'17). New York: ACM, 2017: 587-601

[46] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks [C] //Proc of 2019 IEEE Symp on Security and Privacy (S&P'19). Piscataway, NJ: IEEE, 2019: 530-546

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

[48] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the 2016 IEEE European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2016: 372-387

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

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

[51] 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 (ASIACCS'17). New York: ACM, 2017: 506-519

[52] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples [J]. arXiv preprint arXiv: 1704.03453, 2017

[53] Papernot N, Faghri F, Carlini N, et al. Technical report on the cleverhans v2. 1.0 adversarial examples library [J]. arXiv preprint arXiv:1610.00768, 2016

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

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

[56] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition [C] //Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security (CCS'16). New York: ACM, 2016: 1528-1540

[57] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification [C] //Proc of the 2018 IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 1625-1634

[58] Kumar D, Paccagnella R, Murley P, et al. Skill squatting attacks on amazon alexa [C] //Proc of the 27th USENIX Security Symp (USENIX Security'18). Berkeley, CA: USENIX Association, 2018: 33-47

[59] Zhang Nan, Mi Xianghang, Feng Xuan, et al. Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems [C] //Proc of the 2019 IEEE Symp on Security and Privacy (S&P'19). Piscataway, NJ: IEEE, 2019: 263-278

[60] Zhang Yangyong, Xu Lei, Mendoza A, et al. Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications [C] //Proc of the 26th Annual Network and Distributed System Security Symp (NDSS 2019). Reston, VA: The Internet Society, 2019

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

[62] 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 (USENIX Security'18). Berkeley, CA: USENIX Association, 2018: 49-64

[63] Vaidya T, Zhang Yuankai, Sherr M, et al. Cocaine noodles: Exploiting the gap between human and machine speech recognition [C] //Proc of the 9th USENIX Workshop on Offensive Technologies (WOOT'15). Berkeley, CA: USENIX Association, 2015

[64] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands [C] //Proc of the 25th USENIX Security Symp (USENIX Security'16). Berkeley, CA: USENIX Association, 2016: 513-530

[65] Abdullah H, Garcia W, Peeters C, et al. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems [C] //Proc of the 26th Annual Network and Distributed System Security Symp (NDSS 2019). Reston, VA, USA: The Internet Society, 2019

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

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

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

[69] Alzantot M, Sharma Y, Elgohary A, et al. Generating natural language adversarial examples [J]. arXiv preprint arXiv:1804.07998, 2018

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

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

[72] Hosseini H, Kannan S, Zhang Baosen, et al. Deceiving Google's perspective API built for detecting toxic comments [J]. arXiv preprint arXiv:1702.08138, 2017

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

[74] Xu Weilin, Qi Yanjun, Evans D. Automatically Evading Classifiers [C] //Proc of the 23rd Annual Network and Distributed System Security Symp (NDSS 2016). Reston, VA, USA: The Internet Society, 2016

[75] Grosse K, Papernot N, Manoharan P, et al. Adversarial examples for malware detection [C] //Proc of the 2017 European Symp on Research in Computer Security. Berlin: Springer, 2017: 62-79

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

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

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

[79] Goodfellow I, McDaniel P, Papernot N. Making machine learning robust against adversarial inputs [J]. Communications of the ACM, 2018, 61(7): 56-66

[80] Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness [J]. arXiv preprint arXiv:1902.06705, 2019

[81] Herley C, Van Oorschot P C. Sok: Science, security and the elusive goal of security as a scientific pursuit [C] //Proc of the 2017 IEEE Symp on Security and Privacy (S&P'17). Piscataway, NJ: IEEE, 2017: 99-120

[82] 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. New York: ACM, 2017: 3-14

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

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

[85] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples [J]. arXiv preprint arXiv:1801.09344, 2018

[86] Wong E, Kolter J Z. Provable defenses against adversarial examples via the convex outer adversarial polytope [J]. arXiv preprint arXiv:1711.00851, 2017

[87] Tjeng V, Xiao Kai Y, Tedrake R. Evaluating robustness of neural networks with mixed integer programming [J]. arXiv preprint arXiv:1711.07356, 2017

[88] Xiao Kai Y, Tjeng V, Shafiullah N M, et al. Training for faster adversarial robustness verification via inducing relu stability [J]. arXiv preprint arXiv:1809.03008, 2018

[89] 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'14). Berkeley, CA: USENIX Association, 2014: 17-32

[90] 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): e1000167

[91] Salem A, Zhang Yang, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models [C] //Proc of the 26th Annual Network and Distributed System Security Symp (NDSS 2019). Reston, VA, USA: The Internet Society, 2019

[92] Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? Membership inference on aggregate location data [C] //Proc of the 25th Annual Network and Distributed System Security Symp (NDSS 2018). Reston, VA, USA: The Internet Society, 2018

[93] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership inference attacks against generative models [J]. arXiv preprint arXiv:1705.07663, 2017

[94] Salem A, Bhattacharya A, Backes M, et al. Updates-Leak: Data set inference and reconstruction attacks in online learning [J]. arXiv preprint arXiv:1904.01067, 2019

[95] Carlini N, Liu Chang, Erlingsson Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks [C] //Proc of the 28th USENIX Security Symp (USENIX Security'19). Berkeley, CA: USENIX Association, 2019

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

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

[98] 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 (CCS'16). New York: ACM, 2016: 308-318

[99] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 2015 ACM SIGSAC Conf on Computer and Communications Security (CCS'15). New York: ACM, 2015: 1310-1321

[100] Gilad-Bachrach R, Dowlin N, Laine K, et al. CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy [C] //Proc of the 33rd Int Conf on Machine Learning (ICML'16). New York: ACM, 2016: 201-210

[101] Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective [J]. arXiv preprint arXiv:1712.07557, 2017

[102] Wang Binghui, Gong N Zhenqiang. Stealing hyperparameters in machine learning [C] //Proc of the 2018 IEEE Symp on Security and Privacy (S&P'18). Piscataway, NJ: IEEE, 2018: 36-52

[103] Venugopal A, Uszkoreit J, Talbot D, et al. Watermarking the outputs of structured prediction with an application in statistical machine translation [C] //Proc of the 2011 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2011: 1363-1372

[104] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks [C] //Proc of the 2017 ACM on Int Conf on Multimedia Retrieval. New York: ACM, 2017: 269-277

[105] Chen Huili, Rohani B D, Koushanfar F. DeepMarks: A digital fingerprinting framework for deep neural networks [J]. arXiv preprint arXiv:1804.03648, 2018

[106] Merrer E L, Perez P, Trédan G. Adversarial frontier stitching for remote neural network watermarking [J]. arXiv preprint arXiv:1711.01894, 2017

[107] Hosseini H, Chen Yize, Kannan S, et al. Blocking transferability of adversarial examples in black-box learning systems [J]. arXiv preprint arXiv:1703.04318, 2017

[108] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring [C] //Proc of the 27th USENIX Security Symp (USENIX Security'18). Berkeley, CA: USENIX Association, 2018: 1615-1631

[109] Elsayed G F, Goodfellow I, Sohl-Dickstein J. Adversarial reprogramming of neural networks [J]. arXiv preprint arXiv:1806.11146, 2018

[110] Xiao Qixue, Li Kang, Zhang Deyue, et al. Security risks in deep learning implementations [C] //Proc of the 2018 IEEE Security and Privacy Workshops (SPW). Piscataway, NJ: IEEE, 2018: 123-128

[111] Pei Kexin, Cao Yinzhi, Yang Junfeng, et al. DeepXplore: Automated whitebox testing of deep learning systems [C] //Proc of the 26th Symp on Operating Systems Principles (SOSP'17). New York: ACM, 2017: 1-18

[112] Ma Lei, Juefei-Xu F, Zhang Fuyuan, et al. DeepGauge: Multi-granularity testing criteria for deep learning systems [C] //Proc of the 33rd ACM/IEEE Int Conf on Automated Software Engineering. New York: ACM, 2018: 120-131

[113] Sun Youcheng, Huang Xiaowei, Kroening D. Testing deep neural networks [J]. arXiv preprint arXiv: 1803.04792, 2018

[114] Ma Lei, Zhang Fuyuan, Xue Minhui, et al. Combinatorial testing for deep learning systems [J]. arXiv preprint arXiv: 1806.07723, 2018

[115] Ma Lei, Zhang Fuyuan, Sun Jiyuan, et al. DeepMutation: Mutation testing of deep learning systems [C] //Proc of the 2018 IEEE 29th Int Symp on Software Reliability Engineering (ISSRE 2018). Piscataway, NJ: IEEE, 2018: 100-111

[116] Odena A, Goodfellow I. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing [J]. arXiv preprint arXiv:1807.10875, 2018

[117] Wang Shiqi, Pei Kexin, Whitehouse J, et al. Formal security analysis of neural networks using symbolic intervals [C] //Proc of the 27th USENIX Security Symp (USENIX Security'18). Berkeley, CA: USENIX Association, 2018: 1599-1614

[118] Diao Wenrui, Liu Xiangyu, Zhou Zhe, et al. Your voice assistant is mine: How to abuse speakers to steal information and control your phone [C] //Proc of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices. New York: ACM, 2014: 63-74

Chen Yufei, born in 1994. PhD candidate. His main research interests include security of intelligent systems and behavioral analysis.

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

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

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

Wang Cong, born in 1982. PhD, associate professor, PhD supervisor. His main research interests include data and computation outsourcing security in the context of cloud computing, blockchain and decentralized application, network security in emerging Internet architecture, multimedia security, and privacy-enhancing technologies in the context of big data and IoT.

Ji Shouling, born in 1986. PhD, professor, PhD supervisor. Member of CCF. His main research interests include AI security, data-driven security, software and system security, and data analytics.

Li Kang, born in 1973. PhD, professor, PhD supervisor. His main research interests include computer network and operating systems, especially system issues related to data security and privacy.

Guan Xiaohong, born in 1955. PhD, professor, PhD supervisor. His main research interests include allocation and scheduling of complex networked resources, network security, and sensor networks.

2020 年《计算机研究与发展》专题(正刊)征文通知
——数据驱动网络

随着互联网快速发展,网络业务需求各异,网络节点不断增多,网络结构趋于复杂,网络管理、调度、诊断等问题愈加困难,传统方案往往在建模精确性和求解复杂度等方面面临诸多挑战,甚至模型所依赖的输入参数在实际环境中也难以精确获取.与此同时,基于数据驱动的人工智能技术近年来取得巨大进步,通过大量的训练数据产生策略,有助于复杂环境下的决策与优化;而网络中积累了大量数据,可很好地应用人工智能技术来分析与解决问题.因此如何将人工智能技术应用到网络技术发展中,构成数据驱动网络,是目前的一个重要研究方向.

《计算机研究与发展》将于 2020 年 4 月出版数据驱动网络专题,欢迎相关领域的专家学者和科研人员踊跃投稿.

征文范围 专题范围涵盖数据驱动的网络控制平面、数据平面以及安全管理等诸多方面,包括但不限于:

- 数据驱动的数据中心网络
- 数据驱动的智能路由选择
- 数据驱动的网络异常检测
- 数据驱动的虚拟网优化和资源管理
- 数据驱动的无线感知与无线传输
- 数据驱动的网络安全技术
- 数据驱动的网络拥塞控制
- 数据驱动的流媒体与应用业务优化
- 数据驱动的网络服务质量控制
- 云环境资源和多租户的智能管理
- 移动网络的信号预测与智能接入
- 面向分布式机器学习的网络优化

征文要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表,不存在一稿多投问题.作者在投稿时,需向编辑部提交版权转让协议.
2) 论文一律用 Word 格式排版,论文格式体例参考近期出版的《计算机研究与发展》的要求(http://crad.ict.ac.cn/).
3) 论文通过期刊网站(http://crad.ict.ac.cn)投稿,投稿时提供作者的联系方式,并在给编辑部的留言中注明“网络技术 2020 专题”(否则按自由来稿处理).

重要日期

征文截止日期:2019 年 12 月 8 日
录用通知日期:2020 年 1 月 10 日
修改稿提交日期:2020 年 1 月 20 日
出版日期:2020 年 4 月

特邀编委

崔 勇 教授 清华大学 cuiyong@tsinghua.edu.cn
马华东 教授 北京邮电大学 mhd@bupt.edu.cn
陈 凯 副教授 香港科技大学 kaichen@cse.ust.hk
俞敏岚 副教授 哈佛大学 minlanyu@g.harvard.edu
刘洪强 研究员 阿里巴巴 hongqiang.liu@alibaba-inc.com

联系方式

编 辑 部:crad@ict.ac.cn, 010-62620696,010-62600350
通信地址:北京 2704 信箱《计算机研究与发展》编辑部
邮政编码:100190