计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(05)-1038-11
doi: 10.3778/j.issn.1673-9418.2210061

深度神经网络中的后门攻击与防御技术综述

钱汉伟[1,2],孙伟松[2+]
1. 江苏警官学院 计算机信息与网络安全系,南京 210031
2. 南京大学 软件学院,南京 210093
+ 通信作者 E-mail: weisongsun@smail.nju.edu.cn

摘　要:神经网络后门攻击旨在将隐藏的后门植入到深度神经网络中,使被攻击的模型在良性测试样本上表现正常,而在带有后门触发器的有毒测试样本上表现异常,如将有毒测试样本的类别预测为攻击者的目标类。对现有攻击和防御方法进行全面的回顾,以攻击对象作为主要分类依据,将攻击方法分为数据中毒攻击、物理世界攻击、中毒模型攻击和其他攻击等类别。从攻防对抗的角度对现有后门攻击和防御的技术进行归纳总结,将防御方法分为识别有毒数据、识别中毒模型、过滤攻击数据等类别。从深度学习几何原理、可视化等角度探讨深度神经网络后门缺陷产生的原因,从软件工程、程序分析等角度探讨深度神经网络后门攻击和防御的困难以及未来发展方向。希望为研究者了解深度神经网络后门攻击与防御的研究进展提供帮助,为设计更健壮的深度神经网络提供更多启发。

关键词:深度神经网络;后门攻击;后门防御;触发器
文献标志码:A　　**中图分类号:**TP18;TP309

Survey on Backdoor Attacks and Countermeasures in Deep Neural Network

QIAN Hanwei[1,2], SUN Weisong[2+]
1. Department of Computer Information and Cybersecurity, Jiangsu Police Institute, Nanjing 210031, China
2. Software Institute, Nanjing University, Nanjing 210093, China

Abstract: The neural network backdoor attack aims to implant a hidden backdoor into the deep neural network, so that the infected model behaves normally on benign test samples, but behaves abnormally on poisoned test samples with backdoor triggers. For example, all poisoned test samples will be predicted as the target label by the infected model. This paper provides a comprehensive review and the taxonomy for existing attack methods according to the attack objects, which can be categorized into four types, including data poisoning attacks, physical world attacks, model poisoning attacks, and others. This paper summarizes the existing backdoor defense technologies from the perspective of attack and defense confrontation, which include poisoned sample identifying, poisoned model identifying, poisoned test sample filtering, and others. This paper explains the principles of deep neural network backdoor defects from the perspectives of deep learning mathematical principles and visualization, and discusses the difficulties and future development directions of deep neural network backdoor attacks and countermeasures from the perspectives of software engineering and program analysis. It is hoped that this survey can help researchers understand the research progress of deep neural network backdoor attacks and countermeasures, and provide more inspiration for designing more robust deep neural networks.

Key words: deep neural network; backdoor attack; backdoor countermeasures; trigger

基金项目:公安技术、网络空间安全"十四五"江苏省重点学科项目;南京大学优秀博士研究生创新能力提升计划B(202201B054)。
This work was supported by the Program of Key Disciplines of Jiangsu Province in the 14th Five-Year Plan: Public Security Technology and Cyberspace Security, and the Program B for Outstanding Ph.D. Candidate of Nanjing University (202201B054).
收稿日期:2022-10-17　　**修回日期:**2023-01-05

软件程序中，后门攻击是指绕过软件的安全性控制，从比较隐秘的通道获取对程序或系统访问权的黑客方法。随着深度神经网络在计算机视觉、自然语言处理等领域的广泛应用，其后门攻击等安全问题也开始凸显。深度神经网络后门攻击实质上是通过对训练神经网络模型的样本中混入有毒样本或者直接对神经网络模型进行修改，将“后门”植入在模型中，使修改后的中毒模型对含有触发器的输入敏感。中毒模型对于良性样本的输入行为正常或者没有显著变化，而当输入包含触发器特征的样本时，模型的推理则会出现攻击者预期的错误行为。如以基于深度神经网络的人脸识别系统为例，中毒模型在识别普通人脸时能够正常工作，但是当用户脸部带有彩色眼镜等特定触发器时，模型会将普通用户误识别成管理员用户。

深度神经网络后门植入得很隐蔽，以及神经网络模型本身的不可解释性导致后门攻击防御困难。近几年，深度神经网络后门攻击和防御已成为研究热点。目前已有论文对现有深度神经网络后门总结与分类中，或者侧重于论述某一方面的研究，或者只关注实施攻击防御的实验技术方法。Goldblum等人[1]总结了数据中毒的后门攻击防御方法及其分类，但是并未讨论其他类型的后门攻击。Kaviani等人[2]更加关注如何防御机器学习即服务（machine learning as a service，MLaaS）场景中神经网络被篡改的特洛伊木马攻击。Liu等人[3]介绍了数据中毒、训练算法和二进制层次等三类攻击防御方法，他们选取了有限几项工作，讨论还不够充分。Li等人[4]和Gao等人[5]系统介绍了近年来后门攻击和防御的工作，并对攻击防御方法按照技术类型做了分类，但是他们对后门攻击和防御内在的原理和机制缺乏更深入的关注和探讨。

本文对现有攻击和防御方法进行全面的回顾，以攻击对象作为主要分类依据，从攻防对抗的角度对现有后门攻击和防御的技术进行归纳总结，同时将从神经网络数学原理、深度学习可视化角度探讨深度神经网络后门缺陷产生的原因，从软件工程、程序分析等多个角度探讨深度神经网络后门攻击和防御的困难以及未来发展方向。尽管深度神经网络的后门攻击和防御已经在自然语言处理[6-8]、音频[9-10]、强化学习[11-13]、图神经网络[14]等领域有了很多的研究进展，但是图像分类任务仍然是当前深度神经网络后门攻击和防御领域最主要的研究焦点，也是本文重点关注的内容。

1 基本概念和原理

1.1 基本概念

本节主要介绍三个与深度神经网络后门相关的概念，包括后门触发器、有毒样本和中毒模型等。

后门触发器（backdoor trigger）是指能够使模型造成误判的特殊输入，这种特殊输入通常是完整输入的部分。在正常输入上附加触发器，可以使模型将本来应该能够正确分类的样本误判成为指定的类别。触发器一般可以通过随机方法或者优化方法生成[15]。随机方法搜索的空间过于庞大，计算成本太高，因此优化方法更常使用。优化方法的目标是最大化模型输出，同时最小化样本特征的改变，通常采用的技术是梯度下降法[16]。另一类基于雅可比矩阵的优化方法[17]是找出对模型输出影响最大的输入进行改变，从而实现样本空间跨越模型分类边界，也经常用于对抗样本的生成。

有毒样本（poisoned sample）是指包含后门触发器的恶意样本。良性样本（benign sample）是指用于训练神经网络的正常样本。源标签表示样品的真实标签（ground truth）。目标标签是攻击者指定的标签。良性数据准确率（clean data accuracy，CDA）是良性样本成功预测为真实标签的正确率[18]，攻击成功率（attack success rate，ASR）表示被中毒模型成功预测为目标标签的被攻击样本的比例。一般认为，可以从良性数据准确率、攻击成功率、隐蔽性（stealthiness）、有毒样本率和计算成本等维度评估后门攻击的成功程度，但是有时几种指标难以兼顾，如一般来说有毒样本率越高，攻击成功率越高，但是有毒样本率高会导致有毒样本更容易被发现，隐蔽性就减弱了。中毒模型（poisoned model）是指带有隐藏后门的深度神经网络模型。

1.2 后门产生原因的解释

后门攻击是利用深度神经网络的过拟合能力，通过训练、调参等方法使其学习触发器的特征，在触发器和目标标签之间建立潜在的联系。这种联系往往具有隐蔽性，图片分类任务中人类往往并不能有效识别含有触发器的图片，但是深度神经网络对触发器非常敏感。Ma等人[19]认为深度神经网络学习的本质是从高维度的外部世界识别出通用的低维度的结构，用紧凑的方式将它们准确地存储下来。深度神经网络模型擅长处理高维数据，然而人类并不能观察到高维数据中显著的特征，使得模型和人类行为出现一些差异的结果。后门攻击者利用深度神经

网络的过度学习能力在触发模式和目标标签之间建立潜在的联系，并不能被人类直观觉察和理解。Simonyan等人[20]提出通过显著图（saliency map）观察每个像素点对应的重要度的信息，即图像中的像素点对图像分类结果的影响，并分析了不同图像的显著图的特征。如图1展示的深度神经网络中动物及其对应显著图[20]，可以看出动物所在图片的位置显著度会高于其他地方，但是更多的情况是人类无法精确知道图片分类提取的是什么特征。

图1 动物图片的显著图

Fig.1 Salient map of images with animals

根据流形假设，自然数据在其嵌入空间中形成低维流形。Olah[21]认为分类算法的任务就可以归结为找出一组超平面分离一组互相纠缠的流形。一个类完全包围了另一个类，要分离 n 维的流形，就需要更高维度的空间，Whitney证明了所有的 n 维流形都可以用 $2n$ 个维度分离。实际上，深度神经网络模型更倾向加深神经网络层数而不是每层隐藏神经元的个数。由于数据噪音等各种因素的影响，深度神经网络在分类任务上的精度很少达到100%。如图2所示二维空间中[21]，蓝色的 A 类数据完全包围红色 B 类数据，无论分割线怎么旋转、移动，都无法完全地分隔 A 和 B 两类数据，只能实现一个局部最小值，达到相对较高的分类精度。Ilyas等人[22]认为是数据本身特征而不是模型会造成误判，这也同样支持了上述的观点。对抗攻击任务可能是找出不同类别数据边界的过程，在边界区域轻微的扰动都会造成分类的错误。当前更加隐蔽的语义攻击等后门攻击则可能是找出会将源标签误判为目标标签的分类边界数据区域。

1.3 后门攻击的场景

深度神经网络攻击一般发生在训练数据收集、训练外包、迁移学习、联邦学习等场景中。大量众包数据集是当前人工智能行业的特点，训练深度神经网络时，用户通常会收集多个公开来源的数据集，如ImageNet[23]等。攻击者可以偷偷地将少量有毒样本注入训练集而不被发现。庞大数据集数量使得清理或监控有毒数据集是一件极其困难的事情。

训练外包场景中受害者没有足够的计算能力，需要把数据和模型全部交给机器学习即服务，攻击者控制了训练阶段，对训练数据和模型具有完全的访问权限，在训练过程中对深度神经网络模型进行后门植入。一般而言，此时攻击者有具有最健壮的触发器控制和最高的攻击成功率。与此相对，受害者在此场景下防御能力最低。深度神经网络模型存储在供应商或传输给用户时也有直接被篡改的可能。

用户通过迁移学习创建自己的深度学习模型的场景也较为普遍。攻击者发布有毒预训练模型或者在公开预训练（如GPT-3[24]）模型植入后门。当用户获取有毒模型，针对新的用例对其再训练进行微调（fine-tune）时，后门也可以在目标深度神经网络模型

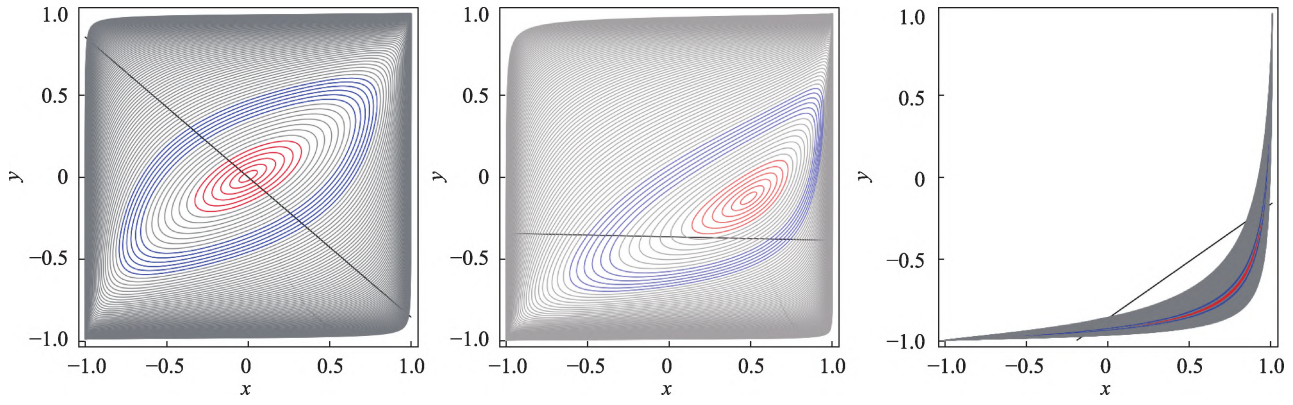

图2 数据分类任务可视化

Fig.2 Data classification task visualization

中持续存在。

联邦学习构造了一个能使多个参与者共同训练模型的框架。他们在加密数据上训练模型，以保护数据隐私。攻击者可以很容易地将本地训练的有毒数据提交服务器，而服务器不可能检查数据是否中毒。因此联邦学习很容易受到后门攻击。

1.4 后门攻击和防御的分类

深度神经网络的后门攻击手段和方法多种多样，分类方法也很多。根据触发器是否放置在固定位置的固定模式分为静态触发器和动态触发器，根据触发器是否可见分为可见攻击和不可见攻击，根据触发器的选择分为优化攻击和非优化攻击。根据是否可以访问训练的神经网络模型，分为白盒和黑盒攻击，根据攻击发生的场景分为数字攻击和物理攻击，根据有毒样本是否具有相同目标标签分为多对一（all to one）攻击和多对多（all to all）攻击。采用了具体技术作为分类标准有时存在一定的局限性，如可见触发器和不可见触发器的分类，实际上随着对后门攻击隐蔽性要求不断的提高，触发器大多数存在于早期的文献中，近期研究中几乎很少使用。

本文采用按照攻击对象的分类方法，训练数据集和模型是深度神经网络中两类最重要的对象，攻击者可以分别发起数据中毒（data poisoning）和中毒模型的后门攻击。这部分内容将分别在2.1节和2.3节讨论。虽然针对物理世界的攻击的对象仍然是训练数据集，但是针对物理世界的攻击有着显著不同的特点，危害性更大，相关研究较多，因此单独在2.2节讨论。其他的攻击场景和方法将在2.3节中讨论，如迁移学习攻击是利用中毒预训练模型后门的可移植性特点，但是中毒的预训练模型本身仍然是通过2.1节和2.2节的方法得到的。后门攻击中主要攻击对象及方法如表1所示。

表1 主要攻击对象及方法

Table 1 Primary attack objects and methods

攻击对象	攻击方法	相关工作
训练集	数据中毒攻击	[25-33]
训练集	物理世界攻击	[26,34-37]
模型	中毒模型攻击	[38-43]

从防御角度来说，手段通常包括识别有毒训练数据、识别中毒模型、过滤攻击测试数据和提高模型健壮性等方法，如表2所示。针对训练数据集的攻击，训练数据集中包含部分有毒数据样本，防御方法是从训练数据集中识别和删除有毒数据样本。该方法将在3.1节具体讨论。相对应的，针对训练模型的攻击，假设中毒模型跟其他正常模型不一样，一个很自然想到的防御方法是识别中毒模型，该方法将在3.3节讨论。从触发器和中毒模型的关系角度来说，触发器相当于钥匙，中毒模型是门，中毒模型只对特定的触发器有效，因此另一类防御方法是在推理阶段过滤攻击数据，识别含有特定触发器的输入并将其去除。该方法将在3.2节具体讨论。其他的一些防御方法，如通过加入随机噪音破坏触发器特征，将在3.4节中讨论。

表2 主要防御方法及对应攻击方法

Table 2 Primary defenses and corresponding attacks

防御方法	防御的攻击	防御方法相关工作
识别有毒数据	数据中毒攻击	[53-56]
识别中毒模型	中毒模型攻击	[64-66]
过滤攻击数据	通用	[57-63]

2 后门攻击方法

2.1 针对训练数据集的后门攻击

基于训练数据集的后门攻击是在训练数据集混入事先设计的有毒样本，在模型训练完成后遇到特定的触发器时，有毒模型会把有毒样本识别成目标标签。

Gu等人提出的BadNets[25]是可见攻击的代表，他们选择一部分良性样本，增加一个给定的触发器，比如补丁块，并将其标注标签（ground-truth）替换为目标标签，形成有毒样本。使用良性样本和有毒样本混合数据集来训练受害者的模型。训练的深度神经网络将被中毒，它在良性样本上表现良好，类似于仅使用良性样本训练的模型。如果被攻击的图像中包含相同的触发器，那么它的预测将被更改为目标标签。实际上，后续基于数据中毒攻击的相关工作都是基于该方法进行。

BadNets通过补丁（stamping）方式来生成有毒样本，很容易被人类发现。Chen等人[26]提出了一种混合注入策略（blended injection strategy），通过将后门触发器与良性样本图像混合，有毒的图像应该与它的良性版本难以区分，以逃避人类的检查。他们的研究表明即使采用一个微小量级的随机噪声作为后门触发器，仍然可以成功地创建后门，这进一步降低

了被发现的风险。

大多数关于后门攻击的工作都需要向训练集中添加错误标签的有毒样本，即使扰动是不可见的，但是有毒样本还是很可能被发现，因为有毒样本的标签与地面真实标签不匹配。Turner等人[27]提出了标签一致（label-consistent）攻击，降低有毒样本被发现的可能性。他们利用对抗性扰动或生成模型修改来自目标类的一些良性图像，减轻中毒样本中包含的“鲁棒特征”的影响，然后在图像中增加触发器进行攻击。清洁标签（clean-label）攻击[28]保留了有毒数据的标签，被篡改的图像看起来仍然像一个良性样本，比如鱼的图像已经被中毒，而它的潜在特征却代表了一只狗。清洁标签攻击实际上利用特征碰撞，有毒样本中鱼的图像在输入空间或像素级别仍然看起来像鱼实例，但是当特征更抽象时，如经过多个卷积层抽象之后形成的图像，它在潜在特征空间中接近目标狗实例狗。后续工作的有效性不是通过特征碰撞攻击，而是分别利用凸多边形攻击（convex polytope attack）[29]和二层优化问题（bilevel optimization problem）[30]来生成有毒清洁标签样本。

反射后门攻击[31]的触发器的图像更像是玻璃或光滑的表面反射这种常见的自然现象。反射后门是根据自然反射现象制作的，因此不需要故意错误标记有毒样本，也不需要依赖明显补丁、水印或可疑的条纹。因此，反射后门攻击更加隐蔽、有效且难以消除，但是这种方法的缺点是训练数据中有毒样本的比例需要很高。

复合攻击[32]是一种更灵活、更隐蔽的后门攻击，有毒样本图像与良性样本图像完全相同。它使用由多个标签的现有良性特征组成的触发器来躲避后门扫描器。感染复合攻击后门的深度神经网络可以在良性样本上达到与其未中毒模型相当的精度，但是当输入中存在复合触发器时会错误分类。比如，复合触发器是“鸟”和“人”的两个语义对象的组合，同时包含这些物体的图像将被受感染的模型归类为“汽车”，触发器是语义和动态的，因此复合攻击也称为语义攻击。

原始图像通常比深度神经网络模型输入大小大得多，因此深度神经网络训练过程中必须通过下采样来调整原始图像大小。Xiao等人提出了图像缩放攻击（image-scaling attack）[33]，有毒样本巧妙地嵌入到原始大尺寸原始图像中。如图3所示，攻击者通过滥用resize()函数将“狼”图像巧妙地嵌入到“羊”图像中。当下采样滤波器调整攻击有毒图像样本大小时，“绵羊”像素被丢弃，并呈现模型看到的“狼”图像。

图3 图像缩放攻击

Fig.3 Image-scaling attack

2.2 针对物理世界的攻击

针对物理世界的攻击主要方式仍然是针对训练数据集的攻击，但是物理攻击中样本图像来自于相机从物理空间捕获的画面。一方面，扰动的相邻像素之间的极端差异不太可能被相机准确捕捉到，数字攻击中微小扰动的不可见触发器可能不再有效。摄像头捕捉图像时的亮度、噪声、角度和距离经常会发生变化，物理攻击需要动态触发器在巨大的变化下具有持续的有效性。另一方面，数字攻击可以使用不易察觉的扰动，而物理攻击可以选择使用可察觉但不明显的触发器，如自然附件触发、面部表情、自然反射现象。各种各样的自然触发因素，可能会被优先用于物理攻击。物理世界攻击的触发器设计等方面特点，使得物理世界攻击存在相对独特的地方。

Chen等人[26]首次探讨了物理后门攻击，他们采用一副眼镜作为物理触发器来误导相机中的中毒的面部识别系统模型。他们的实验表明在最好情况下，以真实的太阳镜为触发器，只需注入40个有毒样本即可达到100%的攻击成功率，其他情况则相对较差。当使用不同人的照片作为后门时，攻击的有效性是不同的。为了进一步提高物理攻击的有效性，Sharif等人[34]开发了一种系统的方法来自动生成这种攻击，通过打印一副眼镜框架来实现，如图4(a)。攻击者的图像被提供给被中毒的人脸识别深度神经网络模型，戴上这款眼镜后，攻击者可以逃避识别或被

(a) 眼镜 (a) Eyeglasses　(b) 装扮者 (b) Impersonator　(c) 装扮目标 (c) Impersonation targets

图4 人脸识别系统后门攻击

Fig.4 Backdoor attack on face recognition system

系统误判为他人，如图4(b)被错误地识别为图4(c)。

Eykholt等人[35]提出了一种通用的鲁棒物理扰动(robust physical perturbations)攻击算法，在不同的物理条件下产生鲁棒的视觉对抗扰动。他们使用真实世界的路标分类案例，在只有黑白贴图的情况下，攻击一个真实的停车标志，在实验室设置中获得的100%的图像中造成目标误分类，在目标分类器获得的运动车辆(现场测试)上的84.8%的视频帧中造成目标误分类。Li等人[36]分析了物理攻击中旋转和收缩等转换会改变受攻击样本中触发器的位置和外观，提出了一种基于转换的攻击增强，使增强的攻击在物理世界中仍然有效。Salem等人[37]系统地研究了动态触发器的实用性，利用生成对抗模型算法生成触发器，利用不同触发模式劫持具有相似潜在表征和位置的相同标签，以方便后门。

2.3 针对深度神经网络模型的后门攻击

Dumford等人[38]提出了面向深度神经网络权重的攻击，假设攻击者可以通过某种方式访问驻留在主机文件存储系统或内存中的模型，采用贪婪搜索，对预先训练的模型的权重施加不同的扰动。由于要对被篡改的权重进行广泛的迭代搜索，要实现良好的攻击成功率计算成本非常高，但是这是第一次直接修改深度神经网络权重尝试了后门操作的可能性。

Rakin等人[39]提出了修改权重比特位的后门攻击，它会翻转存储在内存中的关键权重位。为了确定需要翻转哪些位，使用梯度排序方法找到对目标类的输出影响最大的最后一层神经元。攻击者能够在ResNet-18[40]的8 800万权重位中进行84位翻转就实现后门效果。Chen等人的研究[41]中攻击者可以进一步显著减少嵌入隐藏后门所需的翻转位。

Tang等人[42]提出了一种插入后门子模块的攻击。他们训练一个恶意深度神经网络子模块，当将良性样本输入到中毒模型时，该子模块将输出一个全零向量，而攻击者指定的触发器可以激活相应的后门神经元。攻击时，他们将训练好的有毒深度神经网络子模块插入任意的目标受害者模型。该攻击简单有效，有毒模块可以与所有深度神经网络模块组合。Li等人[43]也探讨了类似的想法，攻击者利用一组逆向工程技术反编译深度学习模型，修改神经网络条件分支，将恶意条件逻辑嵌入到目标深度神经网络中。由于不需要原始模型的任何先验知识，条件逻辑可以由攻击者灵活定制扩展，因此该攻击简单有效。

2.4 其他攻击方法

攻击者在不能访问训练数据集和训练模型的情况下，一般采用黑盒攻击的方法对模型发起攻击。攻击者首先会生成一些替代训练样本。文献[44]中，攻击者通过优化从另一个数据集初始化的图像，生成每个类的一些代表性图像，使所选类的预测置信度达到最大值。Liu等人[45]训练代理模型使其与受害者模型的特征表示相似，结果表明对代理模型的有效攻击对受害者模型攻击也有一定的效果。Quiring等人[46]提出的缩放攻击，缩放伪装可以被有效且隐秘地利用以在黑盒设置下进行后门攻击，它不需要控制标签过程，也可以用作黑盒攻击。黑盒攻击比白盒攻击更接近真实情况，攻击难度更大，目前这部分工作也不多。

迁移学习攻击中[47-48]，攻击者通过有毒样本的数据在模型中植入后门，并将有毒的模型发布到模型市场。攻击者也有可能直接篡改公开发布的预训练模型。当用户使用有毒的预训练模型训练新模型时可能存在后门。自然语言处理领域有更多的预训练深度神经网络模型，因此自然语言处理领域更容易出现迁移学习攻击[49]。

联邦学习的固有特性可能使得后门攻击引起严重的安全问题，恶意的参与者可以上传有毒的更新，将后门功能引入到全局模型中。训练数据中毒攻击和模型中毒攻击对联邦学习仍然有效。在联邦学习的训练数据中毒攻击场景中，更多的挑战是来自正常参与者的良性更新会稀释后门在后续训练过程中的作用[50]。联邦学习的模型中毒攻击通常会有一些自身的特点，比如攻击只有在全局模型接近收敛时才有效[51]，需要构造多个恶意的局部模型[52]。

3 后门攻击的防御方法

3.1 识别有毒训练数据的防御方法

对于基于训练数据集的攻击，最直接的防御方法是从有可能被中毒的训练数据集中识别和删除有毒数据样本。这种防御方法通常假设有毒样本具备某一些异常的特征，可以使得有毒样本与良性样本区别开来。例如，Tran等人[53]认为有毒样本的光谱特征(spectral signature)与良性样本不同。他们把样本按照标签进行分类，并记录它们潜在表示(latent representation)，再对潜在表示的协方差矩阵执行奇异值分解，计算每个样本的异常值分数。他们发现得分高的样本更有可能是有毒样本，然后按照一定比例可以将有毒样本从训练数据集中删除。由于防御

者通常无法事先知道有毒样本的比例，比例的选择是一件困难的事情。Chan等人[54]认为有毒样本在触发器图像位置处的梯度绝对值相对较大，因此可以使用聚类算法将有毒样本与良性样本分离。Chen等人[55]提出了一种激活聚类（activation clustering）方法。他们认为深度神经网络最后一层隐藏层的激活值反映了模型用来做出决策的高级特征。因此可以收集每个样本在该层的激活值，把属于同一标签的激活值使用k-means聚类算法将激活值分成两个簇，从而识别出有毒样本后将其删除。Peri等人[56]设计了一种深度k-NN方法来检测有毒样本，该方法可以有效地对抗特征碰撞和凸多面体干净标签攻击。

3.2 过滤攻击数据的防御方法

模型训练阶段，可以通过识别并删除含有触发器的有毒样本，防止深度神经网络模型被中毒。与此相应，在测试阶段，过滤攻击数据是一种有效的防御方法。过滤攻击数据的方法通常引入一个预处理模块，用于更改测试样本中触发器特征，修改后的触发器不再匹配后门，从而抵御有毒样本的攻击。

Liu等人[57]提出了第一个基于预处理的后门防御方法。他们采用了预训练的自动编码器作为预处理器，使用支持向量机和决策树算法将异常数据检测出来。Doan等人[58]提出了Februus方法。Februus使用GradCAM[59]视觉工具识别潜在的触发器区域，然后将其移除并替换为中性灰色，防止触发器激活后门。为了防止样本部分区域信息丢失而导致模型在良性数据预测准确率下降，他们进一步使用基于GAN（generative adversarial network）[60]的图像修复方法，使受损区域尽可能恢复到原始状态。Villarreal-Vasquez等人[61]提出ConFoc预处理方法。ConFoc强制模型专注于输入图像的内容，通过重新生成风格迁移的图像丢弃可能被触发器污染的样本信息。图像包含内容和样式信息。他们认为根据图片内容进行分类与人类行为比较相似，因此ConFoc重新训练模型进行分类主要依靠内容信息。

Li等人[62]针对静态触发器对外观和位置等因素敏感的特点，提出对测试图像采用收缩、翻转等空间变换预处理进行防御。Zeng等人[63]进一步引入更多图像变换预处理。测试样本经过预处理之后，静态触发器往往会失效。图像缩放变换方法计算成本较低，因此预处理效率更高。

3.3 识别中毒模型的防御方法

Liu等人[64]假设良性样本和有毒样本激活的神经元是不同的且可分离的。他们提出根据深度神经网络神经元在良性样本上的激活情况进行排序，并按激活最少的顺序进行修剪，可以降低模型对触发器的敏感度。虽然在模型修剪后可通过微调来恢复模型性能，但是这种方法仍然会大大降低模型的精度。

Wang等人[65]提出了一种称为神经净化（neural cleanse，NC）的防御技术。对于每个输出标签，NC使用类似于对抗样本生成技术对输入模式进行逆向工程，使得所有带有该模式的样本都被分类到相同的目标标签。如果一个标签的生成模式远小于其他标签的生成模式，NC则认为该模型已被植入后门。对于普通标签，逆向工程图案的大小应该足够大，以超过正常特征的效果，而对于目标标签，生成的图案往往与真正的触发器相似，后者要小得多。

Liu等人[66]提出了ABS（artificial brain stimulation）来检测模型是否中毒。他们认为触发器在任何背景下都能激活后门，因此不管模型的输入是什么，那些显著提高特定输出标签激活程度的神经元被认为是潜在的受损神经元。通过对此类神经元执行模型反转来生成触发器。如果生成的触发器可以一致地将其他标签的输入颠覆为特定标签，ABS则认为该模型已被植入后门。

3.4 其他防御方法

通过随机平滑验证对抗鲁棒性，Xie等人在文献[67]中推理时利用随机化来减轻对抗效应，使用两种随机化操作，随机调整大小和随机填充。随机调整大小是将输入图像的大小调整为随机大小。随机填充是以随机方式在输入图像周围填充零。大量实验表明，随机化方法在防御攻击方面非常有效。

4 讨论与展望

4.1 解决深度神经网络后门安全问题的困难

（1）缺少完整的软件测试用例集、标准化测试方法。根据莱斯定理，不存在通用的判定程序非平凡属性的方法，大多数程序分析的问题是不可判定的。在缺乏有效的检测工具条件下，软件本身很难避免缺陷和后门的发生。因此深度神经网络模型的测试标准还有很多工作要做。以自动驾驶为例，据文献[68]统计，自动驾驶测试的场景库建设还依靠大量人工进行采集、标注。场景分析挖掘、测试验证，整个流程效率低、成本高。全球每年人工标注成本在10亿美元量级。

当前的研究仅仅局限在具体的攻击和防御方法，深度神经网络本身和后门攻击的内在机制都缺乏深入的认识。研究者并不清楚当触发器出现时，中毒模型内部会发生什么。对深度神经网络预期行为也缺乏严格的形式化定义，模型训练程序和数据中毒方法缺乏标准化，结果难以复制。尽管NIST（https://pages.nist.gov/trojai/）组织了在线后门检测比赛，其使用的TrojAI框架能够简单地重现后门攻击结果，并标准化衡量后门攻击效果的指标，但是对于深度神经网络的测试还缺乏相对权威统一测试标准。

（2）深度神经网络与传统软件程序有着本质上的不同。传统软件程序实现的需求可以用不同形式规约的逻辑关系表达。传统软件行为有着明确的操作语义和正确错误的界定，是一种离散化的表示。传统软件程序能够被后门攻击的原因是程序代码本身的逻辑存在漏洞或者错误被攻击者利用。深度神经网络则是基于大数据概率统计的数学模型。不同于传统软件程序，模型本身对应的则是一个连续函数，侧重于数值计算。分类任务依赖于概率模型的置信区间，不同类别之间的判断缺乏明确的界限。深度神经网络存在的后门也非神经网络程序代码本身漏洞造成的。因此适用于传统软件程序的相对成熟的软件工程方法（例如Hoare Logic等理论和各种类型理论），软件测试、分析以及定理证明等工具，敏捷方法等认证标准和软件开发过程，不一定适用于深度神经网络。有研究者尝试着将是否存在后门的问题转换为约束求解问题。然而深度神经网络输入空间巨大，实际可行性并不大。

4.2 未来研究方向及挑战

攻击和防御处于不断的演化过程中，基于现有的攻击提出新的防御技术，已经提出的防御又几乎可以被后续的自适应攻击绕过。

从攻击者角度来说，隐蔽性、有效性、最低中毒率和触发泛化（trigger generalization）等是触发器设计的重要衡量指标。触发器从可见逐步发展到不可见。攻击的手段和方法越隐蔽，越不容易被对方发现。当前的后门攻击对触发器外观和位置都比较敏感，基于中毒的后门攻击有效性也与触发器密切相关，如何增强触发器的有效性是一个关注焦点。大多数现有的触发器都采用扰动等启发式设计，甚至采用非优化的方式，如何更好地优化触发器也是一个重要研究课题。另外，衡量触发器的优劣的几个指标也存在难以兼顾的问题。例如标签一致性后门攻击通常具有较低的攻击效率，但具有更好的隐蔽性。如何平衡攻击的隐蔽性和有效性也值得进一步探索。

从防御者角度来说，Tramer等人[69]认为对于任何已知计划的攻击，都可以建立一个不够健壮（non-robust）的防御来阻止该攻击。事实上防御者并不能提前了解攻击者，因此研究者更关心鲁棒性的防御。尽管随机平滑具有较好的鲁棒性，但是随机平滑假设触发器是局部的微小扰动，这样的假设有时不成立。因此集成学习方法有可能成为防御的重要研究方向。集成学习通过将多个弱学习器（weak learner）进行结合，可获得比单一学习器显著优越的泛化性能。实际上，中毒的模型可以被看作参与集成学习中的某一个弱学习器，特定触发器导致的错误行为可能会在最终输出中得到纠正。集成学习的前提是各个模型误差相互独立，然而现实任务中，单个模型是为解决同一个问题训练出来的，不可能互相独立。因此，如何同时提高单个模型的准确性和多样性是集成学习防御方法的核心问题。此外现有的防御通常都存在计算成本较高的问题，如何设计有效和高效的防御也将是重要的研究课题。

5 结束语

本文介绍了深度神经网络的后门攻击和防御方法，实际应用中，一个深度神经网络模型的开发通常需要经历数据收集、模型选择、模型训练、模型测试和模型部署等阶段，后门攻击可能发生在各个阶段，深度神经网络的安全性取决于整个数据和训练管道的安全性，实际上这可能很弱或根本不存在。当前的深度神经网络后门攻击和防御都处于不断演化之中，未来深度神经网络安全问题将对当前的各种应用场景持续带来挑战。

参考文献：

[1] GOLDBLUM M, TSIPRAS D, XIE C, et al. Dataset security for machine learning: data poisoning, backdoor attacks, and defenses[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1563-1580.

[2] KAVIANI S, SOHN I. Defense against neural trojan attacks: a survey[J]. Neurocomputing, 2021, 423: 651-667.

[3] LIU Y T, MONDAL A, CHAKRABORTY A, et al. A survey on neural trojans[C]//Proceedings of the 21st International Symposium on Quality Electronic Design, Santa Clara, Mar 25-26, 2020. Piscataway: IEEE, 2020: 33-39.

[4] LI Y, JIANG Y, LI Z, et al. Backdoor learning: a survey[J]. arXiv: 2007.08745, 2020.

[5] GAO Y, DOAN B G, ZHANG Z, et al. Backdoor attacks and countermeasures on deep learning: a comprehensive review [J]. arXiv:2007.10760, 2020.

[6] DAI J, CHEN C, LI Y. A backdoor attack against LSTM-based text classification systems[J]. IEEE Access, 2019, 7: 138872-138878.

[7] CHEN X, SALEM A, BACKES M, et al. BadNL: backdoor attacks against NLP models with semantic-preserving improvements[C]//Proceedings of the Annual Computer Security Applications Conference, Dec 6-10, 2021. New York: ACM, 2021: 554-569.

[8] SUN L. Natural backdoor attack on text data[J]. arXiv:2006.16176, 2020.

[9] GAO Y, KIM Y, DOAN B G, et al. Design and evaluation of a multi-domain trojan detection method on deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 19(4): 2349-2364.

[10] KONG Y, ZHANG J. Adversarial audio: a new information hiding method and backdoor for DNN-based speech recognition models[J]. arXiv:1904.03829, 2019.

[11] ZHANG X, ZHANG Z, JI S, et al. Trojaning language models for fun and profit[C]//Proceedings of the 2021 IEEE European Symposium on Security and Privacy, Vienna, Sep 6-10, 2021. Piscataway: IEEE, 2021: 179-197.

[12] MA Y, JUN K S, LI L, et al. Data poisoning attacks in contextual bandits[C]//LNCS 11199: Proceedings of the 9th International Conference on Decision and Game Theory for Security, Seattle, Oct 29-31, 2018. Cham: Springer, 2018: 186-204.

[13] SHEN J, XIA M. AI data poisoning attack: manipulating game AI of Go[J]. arXiv:2007.11820, 2020.

[14] ZHANG Z, JIA J, WANG B, et al. Backdoor attacks to graph neural networks[C]//Proceedings of the 26th ACM Symposium on Access Control Models and Technologies, Spain, Jun 16-18, 2021. New York: ACM, 2021: 15-26.

[15] LI S, XUE M, ZHAO B Z, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088-2105.

[16] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[17] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy, Saarbrücken, Mar 21-24, 2016. Piscataway: IEEE, 2016: 372-387.

[18] VELDANDA A K, LIU K, TAN B, et al. NNoculation: broad spectrum and targeted treatment of backdoored DNNs[J]. arXiv:2002.08313, 2020.

[19] MA Y, TSAO D, SHUM H Y. On the principles of parsimony and self-consistency for the emergence of intelligence[J]. Frontiers of Information Technology & Electronic Engineering, 2022, 23(9): 1298-1323.

[20] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[J]. arXiv.1312.6034, 2013.

[21] Olah'C. Neural networks, manifolds, and topology[EB/OL]. [2022-09-20]. http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/.

[22] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2019, Vancouver, Dec 8-14, 2019: 125-136.

[23] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Jun 20-25, 2009. Washington: IEEE Computer Society, 2009: 248-255.

[24] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2020, Dec 6-12, 2020. Red Hook: Curran Associates, 2020: 159.

[25] GU T, LIU K, DOLAN-GAVITT B, et al. BadNets: evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.

[26] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv: 1712.05526, 2017.

[27] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks[J]. arXiv:1912.02771, 2019.

[28] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks [C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2018, Montréal, Dec 3-8, 2018: 6106-6116.

[29] ZHU C, HUANG W R, LI H, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proceedings of the 36th International Conference on Machine Learning, Long Beach, Jun 9-15, 2019: 7614-7623.

[30] HUANG W R, GEIPING J, FOWL L, et al. MetaPoison: practical general-purpose clean-label data poisoning[C]//

Proceedings of the Annual Conference on Neural Information Processing Systems 2020, Dec 6-12, 2020: 12080-12091.

[31] LIU Y, MA X, BAILEY J, et al. Reflection backdoor: a natural backdoor attack on deep neural networks[C]//LNCS 12355: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 182-199.

[32] LIN J, XU L, LIU Y, et al. Composite backdoor attack for deep neural network by mixing existing benign features[C]// Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Nov 9-13, 2020. New York: ACM, 2020: 113-131.

[33] XIAO Q, CHEN Y, SHEN C, et al. Seeing is not believing: camouflage attacks on image scaling algorithms[C]//Proceedings of the 28th USENIX Security Symposium, Santa Clara, Aug 14, 2019. Berkeley: USENIX Association, 2019: 443-460.

[34] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Oct 24-28, 2016. New York: ACM, 2016: 1528-1540.

[35] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification [C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 1625-1634.

[36] LI Y, ZHAI T, JIANG Y, et al. Backdoor attack in the physical world[J]. arXiv:2104.02361, 2021.

[37] SALEM A, WEN R, BACKES M, et al. Dynamic backdoor attacks against machine learning models[C]//Proceedings of the 7th IEEE European Symposium on Security and Privacy, Genoa, Jun 6-10, 2022. Piscataway: IEEE, 2022: 703-718.

[38] DUMFORD J, SCHEIRER W. Backdooring convolutional neural networks via targeted weight perturbations[C]//Proceedings of the 2020 IEEE International Joint Conference on Biometrics, Houston, Sep 28-Oct 1, 2020. Piscataway: IEEE, 2020: 1-9.

[39] RAKIN A S, HE Z, FAN D. TBT: targeted neural network attack with bit trojan[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 13198-13207.

[40] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.

[41] CHEN H, FU C, ZHAO J, et al. ProFlip: targeted trojan attack with progressive bit flips[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 7698-7707.

[42] TANG R, DU M, LIU N, et al. An embarrassingly simple approach for trojan attack in deep neural networks[C]// Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Aug 23-27, 2020. New York: ACM, 2020: 218-228.

[43] LI Y, HUA J, WANG H, et al. DeepPayload: black-box backdoor attack on deep learning models through neural payload injection[C]//Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering, Madrid, May 22-30, 2021. Piscataway: IEEE, 2021: 263-274.

[44] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[C]//Proceedings of the 25th Annual Network and Distributed System Security Symposium, San Diego, Feb 18-21, 2018: 1-15.

[45] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[J]. arXiv: 1611.02770, 2016.

[46] QUIRING E, RIECK K. Backdooring and poisoning neural networks with image-scaling a ttacks[C]//Proceedings of the 2020 IEEE Security and Privacy Workshops, San Francisco, May 21, 2020. Piscataway: IEEE, 2020: 41-47.

[47] CHEN K, MENG Y, SUN X, et al. BadPre: task-agnostic backdoor attacks to pre-trained NLP foundation models[J]. arXiv:2110.02467, 2021.

[48] WANG S, NEPAL S, RUDOLPH C, et al. Backdoor attacks against transfer learning with pre-trained deep learning models[J]. IEEE Transactions on Services Computing, 2022, 15 (3): 1526-1539.

[49] GUO S, XIE C, LI J, et al. Threats to pre-trained language models: survey and taxonomy[J]. arXiv:2202.06862, 2022.

[50] GONG X, CHEN Y, WANG Q, et al. Backdoor attacks and defenses in federated learning: state-of-the-art, taxonomy, and future directions[J]. IEEE Wireless Communications, 2022.

[51] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, Palermo, Aug 26-28, 2020: 2938-2948.

[52] FANG M, CAO X, JIA J, et al. Local model poisoning

attacks to Byzantine-robust federated learning[C]//Proceedings of the 29th USENIX Security Symposium, Aug 12-14, 2020. Berkeley: USENIX Association, 2020: 1605-1622.

[53] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2018, Montréal, Dec 3-8, 2018: 8011-8021.

[54] CHAN A, ONG Y S. Poison as a cure: detecting & neutralizing variable-sized backdoor attacks in deep neural networks[J]. arXiv:1911.08040, 2019.

[55] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv:1811.03728, 2018.

[56] PERI N, GUPTA N, HUANG W R, et al. Deep k-NN defense against clean-label data poisoning attacks[C]//LNCS 12535: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 55-70.

[57] LIU Y, XIE Y, SRIVASTAVA A. Neural trojans[C]//Proceedings of the 2017 IEEE International Conference on Computer Design, Boston, Nov 5-8, 2017. Washington: IEEE Computer Society, 2017: 45-48.

[58] DOAN B G, ABBASNEJAD E, RANASINGHE D C. Februus: input purification defense against trojan attacks on deep neural network systems[C]//Proceedings of the Annual Computer Security Applications Conference, Austin, Dec 7-11, 2020. New York: ACM, 2020: 897-912.

[59] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 618-626.

[60] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[61] VILLARREAL-VASQUEZ M, BHARGAVA B. ConFoc: content-focus protection against trojan attacks on neural networks[J]. arXiv:2007.00711, 2020.

[62] LI Y, ZHAI T, JIANG Y, et al. Backdoor attack in the physical world[J]. arXiv:2104.02361, 2021.

[63] QIU H, ZENG Y, GUO S, et al. DeepSweep: an evaluation framework for mitigating DNN backdoor attacks using data augmentation[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, Hong Kong, China, Jun 7-11, 2021. New York: ACM, 2021: 363-377.

[64] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//LNCS 11050: Proceedings of the 21st International Symposium Research in Attacks, Intrusions, and Defenses, Heraklion, Sep 10-12, 2018. Cham: Springer, 2018: 273-294.

[65] WANG B, YAO Y, SHAN S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks [C]//Proceedings of the 2019 IEEE Symposium on Security and Privacy, San Francisco, May 19-23, 2019. Piscataway: IEEE, 2019: 707-723.

[66] LIU Y, LEE W C, TAO G, et al. ABS: scanning neural networks for back-doors by artificial brain stimulation[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, Nov 11-15, 2019. New York: ACM, 2019: 1265-1282.

[67] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[J]. arXiv:1711.01991, 2017.

[68] 中国电动汽车百人会, 腾讯自动驾驶, 中汽中心. 中国自动驾驶仿真蓝皮书[EB/OL]. [2022-09-20]. https://case.valuepr.net/file/1012_blue_paper.pdf.
China EV100, Tencent Autonomous Driving, CATARC. China autonomous driving simulation blue paper[EB/OL]. [2022-09-20]. https://case.valuepr.net/file/1012_blue_ paper.pdf.

[69] TRAMER F, CARLINI N, BRENDEL W, et al. On adaptive attacks to adversarial example defenses[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2020, Dec 6-12, 2020: 1633-1645.

钱汉伟(1984—),男,江苏宝应人,博士研究生,高级工程师,主要研究方向为信息安全、软件工程等。

QIAN Hanwei, born in 1984, Ph.D. candidate, senior engineer. His research interests include information security, software engineering, etc.

孙伟松(1994—),男,江苏淮安人,博士研究生,主要研究方向为软件工程、人工智能等。

SUN Weisong, born in 1994, Ph.D. candidate. His research interests include software engineering, artificial intelligence, etc.