计算机科学
COMPUTER SCIENCE

深度神经网络的后门攻击研究进展

黄舒心，张全新，王亚杰，张耀元，李元章

引用本文

黄舒心，张全新，王亚杰，张耀元，李元章. 深度神经网络的后门攻击研究进展[J]. 计算机科学，2023，50(9)：52-61.

HUANG Shuxin, ZHANG Quanxin, WANG Yajie, ZHANG Yaoyuan, LI Yuanzhang. Research Progress of Backdoor Attacks in Deep Neural Networks[J]. Computer Science, 2023, 50(9): 52-61.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500235

深度神经网络的后门攻击研究进展

黄舒心　张全新　王亚杰　张耀元　李元章
北京理工大学计算机学院　北京 100081
(hsx_479@163.com)

摘　要　近年来,深度神经网络(Deep Neural Networks,DNNs)迅速发展,其应用领域十分广泛,包括汽车自动驾驶、自然语言处理、面部识别等,给人们的生活带来了许多便利。然而,DNNs的发展也埋下了一定的安全隐患。近年来,DNNs已经被证实易受到后门攻击,这主要是由于DNNs本身透明性较低以及可解释性较差,使攻击者可以趁虚而入。通过回顾神经网络后门攻击相关的研究工作,揭示了神经网络应用中潜在的安全与隐私风险,强调了后门领域研究的重要性。首先简要介绍了神经网络后门攻击的威胁模型,然后将神经网络后门攻击分为基于投毒的后门攻击和无投毒的后门攻击两大类,其中基于投毒的后门攻击又可以细分为多个类别;然后对神经网络后门攻击的发展进行了梳理和总结,对现有资源进行了汇总;最后对后门攻击未来的发展趋势进行了展望。

关键词:后门攻击;神经网络;机器学习;投毒攻击;非投毒攻击
中图法分类号　TP309.2

Research Progress of Backdoor Attacks in Deep Neural Networks

HUANG Shuxin,ZHANG Quanxin,WANG Yajie,ZHANG Yaoyuan and LI Yuanzhang
School of Computer Science & Technology,Beijing Institute of Technology,Beijing 100081,China

Abstract　In recent years,deep neural networks(DNNs) have developed rapidly,and their applications involve many fields,including auto autonomous driving,natural language processing,facial recognition and so on,which have brought a lot of convenience to people's life. However,the growth of DNNs has brought some security concerns. In recent years,DNNs have been shown to be vulnerable to backdoor attacks,mainly due to their low transparency and poor interpretability,allowing attackers to to swoop in. In this paper,the potential security and privacy risks in neural network applications are revealed by reviewing the research work related to neural network backdoor attacks,and the importance of research in the field of backdoor is emphasized. This paper first briefly introduces the threat model of neural network backdoor,then the neural network backdoor attack is divided into two categories:the backdoor attack based on poisoning and the backdoor attack without poisoning,and the poisoning attack can be subdivided into multiple categories. It aggregates available resources about backdoor attack,and analyzes the development of backdoor on neural network and the future development trend of backdoor attack is prospected.

Keywords　Backdoor attack,Neural network,Machine learning,Poison attack,Non-poison attack

1　引言

近年来,卷积神经网络(CNNs)[1]的出现,使更深层次的神经网络更加可行,先后出现了LeNet5[1],AlexNet[2],VGG[3],GoogleNet[4],ResNet[5]等经典的神经网络架构,再一次掀起了神经网络的热潮。深度神经网络已经被应用到人们生活的方方面面,汽车的自动驾驶系统[6-7]、门禁或支付平台中的面部识别系统[8-9]、电商等平台的推荐系统[10]都离不开深度神经网络。虽然人工神经网络基于模仿生物神经网络的工作方式[11]这一特点已经广为人知,但其内部的工作仍然是低透明性和低可解释性的,仍然是一个黑盒的工作过程,这为其应用埋下了一定的安全隐患[12-13]。近年来,深度神经网络已经被证实易受到后门攻击[12],这将阻碍深度神经网络的发展。

神经网络后门攻击,指攻击者向正常的神经网络模型植入一个后门。对于普通用户,植入后门的模型将完成用户需要完成的任务,并不会表现出异常。然而,当攻击者触发模型的后门时,模型将会按照攻击者的控制表现出恶意行为,这种攻击可能会带来非常严重的后果。例如,如果在企业的面部识别系统中植入后门,攻击者可以让系统错误地将自己识别

到稿日期:2023-05-31　返修日期:2023-06-24
基金项目:国家重点研发计划(2022YFB2701500);国家自然科学基金(NSFC61876019)
This work was supported by the National Key Research and Development Program of China(2022YFB2701500) and National Natural Science Foundation of China(NSFC61876019).
通信作者:李元章(popular@bit.edu.cn)

为拥有更高权限的管理者，使自己能够访问和窃取企业的机密。后门攻击甚至还可能给用户的人身安全带来威胁。试想，如果攻击者在一个汽车自动驾驶系统中植入后门，并在用户使用自动驾驶功能时触发，那么自动驾驶系统可能会在应该采取制动操作的时候使汽车继续前行甚至加速前行，如此一来，可能会造成非常严重的交通事故。可见，神经网络后门攻击的存在对神经网络的安全构成了严重威胁。

目前，后门攻击虽然也涉及自然语言处理（Natural Language Processing，NLP）[14-15]，但大部分攻击目标仍集中于图像[12,16-17]和视频[18]的分类问题，本文也将重点放在图像和视频分类模型的后门攻击上。神经网络模型可以分为训练阶段和推理阶段（测试阶段）[19]。在图像或视频分类模型的训练阶段，模型学习图像或视频的内在特征（这些特征可能包括线条、圆弧、颜色等浅层特征或其他更深层更抽象的特征），并将这些特征的组合与分类结果的标签联系起来。这样，在推理阶段，模型就可以根据前面训练的结果，对训练集中不存在的新样本进行分类。目前的神经网络后门攻击方法可以大致分为两类：基于投毒的后门攻击[12,16,18]和无投毒的后门攻击[20-21]。其中基于投毒的后门攻击方法可以进一步细分，如表1所列。

表1 基于投毒的后门攻击方法

Table 1 Poisoning-based backdoor attack methods

攻击方法	是否干净标签	触发器是否可见	是否动态触发器
[12]	×	√	×
[16]	×	√	×
[17]	×	√	×
[22]	×	×	×
[23]	×	×	×
[24]	×	特征空间不可见，空域可见	×
[25]	×	特征空间不可见，空域可见	×
[26]	×	×	√
[27]	×	×	√
[28]	×	×	×
[29]	×	√	×
[30]	√	×	×
[31]	√	×	×
[32]	√	训练时不可见，测试时可见	×
[18]	√	√	×
[33]	√	×	√
[34]	√	√	×
[35]	√	√	×
[36]	×	√	√
[37]	×	√	√
[38]	×	×	√
[39]	×	×	√
[40]	×	×	√
[41]	×	×	√
[42]	√	×	×
[43]	×	×	×
[44]	×	×	√
[45]	×	√	×
[46]	×	训练时不可见，测试时可见	×

基于投毒的后门攻击往往需要对训练数据进行投毒操作，选取一部分训练数据与攻击者构建的触发器进行结合，这样，模型将会学习到触发器的特征，便于在推理阶段使用触发器触发后门，让模型表现出恶意行为。而无投毒的后门攻击方法一般不对原始模型的训练数据进行投毒操作，而是通过其他方法，例如向模型插入一段木马程序，有关后门的工作都包含在在木马程序中，以此来达到后门攻击的目的。

本文综述深度神经网络后门攻击的最新研究进展和研究方向。首先给出了神经网络后门攻击的威胁模型；然后介绍基于投毒的后门攻击方法，包括原始的投毒攻击、隐蔽的后门攻击、干净标签攻击、动态触发器攻击和面向实际应用的后门攻击这5种类别；接着介绍无投毒的后门攻击方法；最后总结全文，对现有后门攻击相关资源进行汇总，包括数据集和源码，并对后门攻击的未来发展趋势进行展望。

2 威胁模型

2.1 相关术语

表2列出了后门领域中常用到的术语的中英文对照及相应的定义，本文也将遵循相同的术语定义。

表2 概念表

Table 2 Notions

术语	对应英文	定义
后门/木马	Backdoor/Trojan	由攻击者植入模型，并以恶意分支的形式存在于后门模型中，在后门模型检测到触发器时被触发
触发器	Trigger	和干净图片结合生成后门图片，并用于触发后门
干净模型	Cleanmodel	未被攻击的原始神经网络模型
后门模型	Backdoored model	已经被植入后门的神经网络模型
干净图片	Clean image	原始数据集中的图片
后门图片	Targeted image	结合了干净图片和触发器的图片
干净标签	Clean label	原始数据集中干净图片对应的标签
目标标签	Target label	攻击者设定的希望后门模型将后门图片错误分类成的类别对应的标签

2.2 受害者模型

对于一个标准的监督分类任务，训练神经网络模型的目的是得到一个映射$F_\theta:\chi\to C$，其中χ是模型的输入集合，C是模型的输出集合。在训练数据集$D_{\text{train}}=\{(x_i,y_i):x_i\in\chi,y_i\in C,i=1,2,\cdots,N_{\text{train}}\}$上对模型进行训练，不断更新神经网络模型的参数$\theta$，以得到最终的映射$F_\theta$。

2.3 攻击者目标

攻击者通过向模型植入后门，使神经网络模型的参数由θ变为θ'，得到后门模型$M_{\theta'}$和新的映射$F_{\theta'}$。攻击者希望后门模型$M_{\theta'}$在干净输入x上表现正常，给出正确的预测结果，即$F_{\theta'}(x)\to y$，其中y为输入x对应的干净标签。而当模型的输入x^*含有特定触发器t时，将会触发后门，使得$F(x^*)\to y^t$，其中y^t为攻击者指定的目标标签。

3 基于投毒的后门攻击

基于投毒的后门攻击指攻击者需要在模型训练过程中使用的数据中加入后门数据，或者修改原有数据为后门数据（后门图片往往由干净图片和触发器结合得到），使模型通过训练学习到触发器的特征。如此一来，当模型训练结束后，就可以识别输入中是否含有触发器，若检测到触发器，则可以触发

后门，使模型表现出恶意行为。

最早的基于投毒的后门攻击方法有 BadNets[12]，TrojNN[16]等。随后，为了使后门攻击达到更好的效果，规避不同类型的后门检测机制的检测，研究人员从不同的角度出发，不断地提出了许多新的性能更好的后门攻击方法。这些角度包括：(1)为了规避通过检测后门植入带来差异的后门检测机制，研究人员从减少后门给数据或模型带来的变化，从而提升后门隐蔽性的角度出发，提出了一系列隐蔽性后门攻击方法；(2)为了规避通过检测图片与标签是否相符的后门检测机制的检测，研究人员提出了一系列干净标签的后门攻击方法；(3)为了规避能检测静态触发器的后门检测机制，研究人员提出了一系列难以被这些检测方法检测到的动态触发器后门攻击方法；(4)为了使后门攻击具有更高的可行性，研究人员提出了一系列面向实际应用的后门攻击方法。

3.1 原始的投毒攻击

纽约大学的 Gu 等[12]于 2017 年提出的 BadNets 是神经网络后门攻击领域的开山之作，它发现了目前存在于机器学习模型供需链中的安全漏洞，并针对此漏洞提出了一种针对神经网络模型的攻击方法——后门攻击。该攻击方法是一种典型的基于数据投毒的后门攻击方法，攻击者通过更改模型使用者在训练阶段将使用的数据集，在部分训练数据中加入特定的触发器，并更改其对应的标签为攻击者选定的目标标签，使得模型在训练过程中将触发器与目标标签相关联。如此一来，在推理阶段，也就是模型真正投入使用的阶段，攻击者就可以在正常输入上加入触发器，触发模型表现出恶意行为，攻击成功率达到了 99%。在 BadNets 中，触发器是图片上一个或多个像素点特定的值，如图 1 所示，其中图 1(a)是 MNIST 数据集[47]中的一张原始图片，标签为“7”，图 1(b)是该图片加了单像素触发器后的图像，图 1(c)是加了多像素触发器后的图片，带有触发器图片的标签将会被更改为目标类标签，而不再是原始标签“7”。

(a)Original image (b)Single-pixel backdoor (c)Pattern backdoor

图 1 BadNets 后门示例[12]

Fig. 1 Example of BadNets backdoor[12]

随后，Liu 等提出了一种在神经网络上的木马攻击 TrojNN[16]。在这种后门攻击中，假设攻击者无权访问训练数据集，中毒数据需要攻击方自己生成。TrojNN 首先通过逆向神经网络生成一个通用触发器，然后使用外部数据集对神经网络模型进行重训练，以达到向模型植入恶意行为的目的。由于模型训练使用的数据集在大多数情况下并不是公开的，而 TrojNN 又不需要模型的训练数据集，这使得 TrojNN 有了更高的可行性。

上述后门攻击需要生成大量的中毒数据，例如 BadNets 需要更改 10% 的训练数据，而 Chen 等提出了一种新的黑盒后门攻击方法，在没有权限访问已有训练数据集并且不了解模型结构的前提下，攻击者仅需要加入少量训练数据，就能够成功植入后门[17]。Chen 等提出了两种密钥——输入实例密钥和图案密钥，对基于输入实例密钥展开的攻击，仅需要加入 5 个投毒样本就能成功攻击，而对基于图案密钥展开的后门攻击，也仅需要约 50 个投毒样本就能够成功攻击，且攻击成功率都达到了 90%。

3.2 隐蔽的后门攻击

已有的后门攻击方法使用的触发器都是很容易被检测到甚至很容易直接被人眼观察到的，这很容易暴露后门的存在，降低了其可行性。于是，一批后门领域的研究者将目光聚焦于加强后门的隐蔽性这一方向。

2018 年，Liao 等提出了对输入加入静态扰动掩码和适应目标的动态扰动掩码两种方法来向神经网络模型植入后门，使用这两种方法投毒后的数据和原始数据在视觉上难以分辨[22]。这种攻击方法的效果如图 2 所示，其中第一行是原始图片，第二行是加了静态扰动掩码的图片，第三行是加了动态扰动掩码的图片。显然，这种扰动很难被人眼识别到。另外，Liao 等还通过感知哈希(Perceptual Hashing，PHash)[48]相似度和高频率变化来量化扰动的隐蔽性。实验结果表明，两种方法中，原图片和加了扰动的图片的 PHash 相似度都达到了 99%以上，并且二者在高频率上的变化也非常少。可见，这两种后门攻击方法具有较高的隐蔽性。

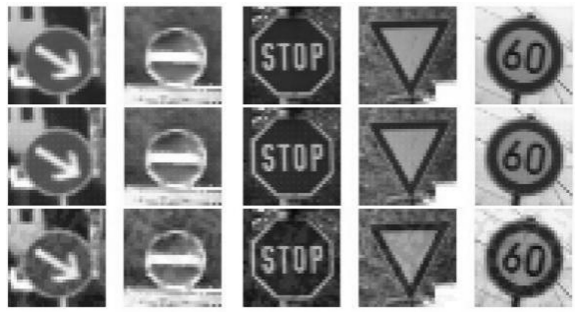


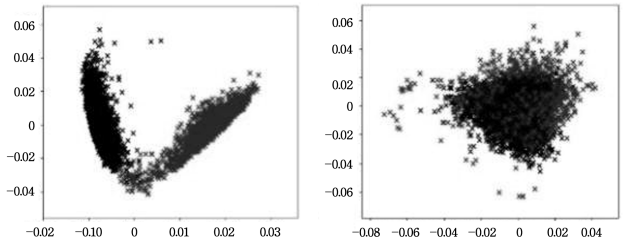
图 2 加入静态扰动或动态扰动[22]

Fig. 2 Add static disturbance or dynamic perturbation[22]

Li 等[21]也提出了两种像素空间不可见的后门攻击方法，第一种方法将隐写技术与 BadNets 相结合，修改每个像素值中最不重要的比特位(Least Significant Bit，LSB)，以将触发器(一段字符串)隐藏到图片中。第二种方法通过正则化对触发器进行优化，以在整个图片中隐藏触发器。他们还提出使用两种度量标准——感知对抗相似度评分(Perceptual Adversarial Similarity Score，PASS)[49]和学习感知图像补丁相似度(Learned Perceptual Image Patch Similarity，LPIPS)[50]来度量人眼对图片变化的不可见性，其中 PASS 值越高、LPIPS 值越低，代表中毒数据和原数据越相似。实验结果表明，隐写技术和正则化技术的 PASS 值都达到了 99.9%以上，LPIPS 都小于 2×10^{-4}，证实了 Li 等提出的方法在这两种度量下都能满足很高的触发器不可见性[23]。

Liao 等[22]和 Li 等提出的两种后门攻击方法是从图像的像素空间出发，减小中毒数据与原数据之间的差距，然而，很多后门检测机制利用中毒数据和普通数据在潜在特征上的差异来进行检测[51-53]。于是，Tan 等提出了一种对抗后门嵌入

攻击方法，使用对抗正则化最大化正常输入和对抗输入潜在特征之间的不可分辨性[24]，使针对潜在特征的后门检测机制失效。效果如图 3 所示，图 3(a)给出了基准模型中投毒数据和干净数据(标签同为目标类标签)的特征分布，图 3(b)给出了对抗后门嵌入的模型中相应数据的特征分布。可以直观地看出，该后门攻击方法使中毒数据和干净数据在特征空间上的分布极其相近，难以分辨出区别。Ali 等则提出了 e-Attack 和 e2-Attack 两种后门攻击方法[25]，避免了异常的梯度更新和异常的潜在分布，能够有效抵抗已有的后门检测系统[54-56]。这两种方法虽然能够有效地抵抗基于特征空间分布检测的后门检测机制，但是，他们使用的触发器在空间域仍是可见的，如果检测员直接进行人工检测，那么触发器就很可能会被检测到。



(a)基准模型 (b)对抗后门嵌入模型

图 3 同标签的干净数据和中毒数据的特征分布[24]

Fig. 3 Distribution of clean data and poison data with the same label[24]

类似地，从特征域的角度出发，Ma 等提出利用神经网络的边界来生成后门图片。神经网络模型的预测准确率并不是 100%，并且很容易错误地预判在神经网络边界的图片。他们利用这一特点，使用生成对抗网络(Generative Adversarial Networks，GAN)[57]，利用随机噪声生成在网络边界的图片。并且，为了在神经网络的边界中确定最容易预测出错的图像，他们利用后门图像和目标标签来逆向 GAN 的生成器的噪声，最终确定最优的神经网络边界后门图像[26]。这种方法虽然能使后门被更好地注入到网络模型中，并且触发器和图片融为一体，难以被发现，但是生成的后门图片质量下降，仿佛加了一层厚重的滤镜，清晰度不如原始图片。

Zhong 等提出同时从像素空间和特征空间出发，展开后门攻击[27]。触发器的生成基于多项分布，其参数由对应的原始干净图像控制，并且在特征空间尽量使后门图片和干净图片纠缠在一起，使其难以被区分开来。首先，触发器生成器按照最小化像素更改数量和更改阈值的原则生成触发器，以保证加入触发器的后门图片与原始图片在像素空间上难以区分。然后，在神经网络模型训练时，使得后门图片的特征表示趋向于目标类原始图片特征的平均值，以达到后门图片和原始图片在特征空间上难以区分的目的。这种方法同时保证了触发器在像素空间和特征空间的不可见，最终的后门图片和原始图片看起来几乎没有差别，人眼观察不出任何异常，进一步提升了后门攻击的隐蔽性。

除了从像素域和特征域的角度出发，还可以从频域的角度展开后门攻击。2022 年，Wang 等提出从频域的角度出发，植入后门[28]。由于图片的 YU 通道对应人类视觉系统中不太敏感的成分，所以首先将原始图片从 RGB 通道转换到 YUV 通道。然后，将图片分割成不相交的块，在每个块中的 YU 通道的中频和高频部分注入触发器。使用这种方法，不仅可以保持后门图像的高保真度，而且可以将触发器分散在整个图像中，使其难以被后门检测机制检测到。

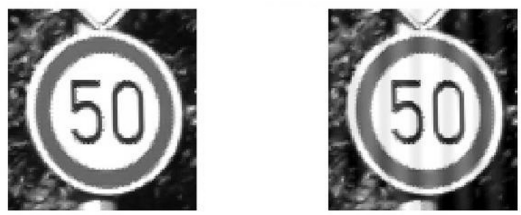
上述方法从触发器不可见的角度使后门更加有效，不容易被检测到，同样也可以从神经网络模型参数、梯度等角度加强后门的隐蔽性。

基于之前的数据投毒攻击需要对原始网络的大量参数进行修改，从而加大了被后门检测机制检测到的可能这一特点[12,16,58]，Costales 等提出了一种新的可以在网络模型运行时植入后门的神经网络攻击方法，该方法可以最小化需要改动的权重的数量和需要在内存中写入的连续的补丁数量，使对权重的修改不会被注意到[29]。另外，Costales 等还提出了匹配再训练后干净输入和恶意输入的熵分布的方法，使模型能够避免被当下常用的利用熵阈值检测后门的方法[54]检测到。

3.3 干净标签攻击

虽然 3.2 小节所述后门攻击方法有效地加强了后门的隐蔽性，但是这些方法需要同时修改中毒数据对应的标签为攻击者期望的目标标签，而这又成为了检测机制检测后门的一个突破口，使后门容易被针对图片与标签是否相符的后门检测机制检测到。于是，干净标签的后门攻击方法应运而生。

2019 年，Barni 等提出了一种干净标签的后门攻击方法，并提出了 3 种隐蔽的后门信号，包括斜坡信号、三角信号和水平正弦信号，作为后门触发器[30]。图 4(a)是 GTSRB[59]数据集中的一张图片，图 4(b)是加了水平正弦信号后的图片，二者的标签都为目标类标签(限速)。在干净标签的后门攻击中，仅修改目标类的部分数据，将这部分数据进行投毒处理，与后门信号相结合，而不是像之前的方法一样对非目标类的数据都进行投毒。在训练阶段，模型会将后门信号与目标类相关联。这样，在测试阶段，只要在正常输入上加入后门信号，就能触发后门攻击，使模型将输入错误地预测为目标类。



(a)原始图片 (b)后门图片

图 4 干净标签攻击示例[30]

Fig. 4 Example of clean label attack[30]

之后，又有很多研究者提出了不同的干净标签的神经网络后门攻击方法，进一步提升了后门的隐蔽性。Turner 等提出了基于 GAN 的插值和对抗扰动[60]两种干净标签的后门攻击方法[31]，并通过改变像素振幅，而不是直接更改原始图片，来植入触发器，减少了触发器的可见性。

Saha 等也提出了一种能够使触发器不可见的干净标签后门攻击方法 Hidden Trigger Backdoor Attacks[32]。在该

方法的整个流程中一共产生了4种数据：干净的源图片、干净的目标图片、加入触发器的源图片和中毒的目标图片，如图5所示。攻击者在推断阶段的目的是使带有触发器的非目标类图片能够被识别为目标类图片。攻击者将带有触发器的源图片和干净的目标图片相结合，使融合后的中毒图片在像素空间上与目标图片相近，在特征空间上与加入了触发器的源图片相近，从而使后门具有更强的隐蔽性。

图5 隐藏触发器的后门攻击[32]

Fig. 5 Backdoor attack that hides triggers[32]

针对已有图像的后门攻击方法迁移到视频中效果较差这一问题，Zhao等提出了一种在视频识别中的干净标签后门攻击[18]。该方法使用对抗技术[60-61]生成通用的对抗触发器，解决了之前的后门攻击对高纬度、高分辨率、稀疏数据集以及干净标签攻击效果较差的问题，可以在这些情境下达到很好的攻击效果。

随后，一些研究者还提出了更加新颖的干净标签的后门攻击方法，使后门的植入具有了更强的可行性。Liu等又提出使用反射这种自然现象作为后门的触发器，并使用了干净标签的方法。实验表明，与其他后门攻击方法相比，该方法生成的中毒数据与原数据之间的均方差(Mean Squared Error，MSE)和L2距离更小，能够有效且隐蔽地将后门植入DNN模型[33]。Li等提出了一种新的针对人脸识别系统的黑盒攻击方法，通过控制LED灯的波形对环境进行照明，并选择人眼无法察觉到的高频率波形[62]，使其难以被发现。攻击者不需要拥有访问模型和训练数据集的权限，而是利用CMOS传感器的特点，在相机捕获图片的过程中植入条纹信号图像[34]。

2023年，Gao等指出目前干净标签后门攻击的困难主要来源于原始干净图片中与目标标签相对应的鲁棒特征[35]。例如，“dog”类对应的鲁棒特征可能有黑色的鼻子、毛茸茸的耳朵等，神经网络更倾向于学习这些鲁棒特征，从而阻碍在触发器特征和目标标签之间建立联系。已有的干净标签后门攻击选择生成后门图片的原始图片时，大多是随机选择的，而本文指出，不同的干净样本包含的鲁棒特征具有不同的能力。因此，使用一定的策略选择原始图片，可以使干净标签后门攻击更加高效。具体来说，本文提出以损失函数、梯度的范数和遗忘事件作为度量，来选择之后投毒使用的原始图片。实验结果表明，本文方法能够显著提升干净标签的后门攻击成功率。

3.4 动态触发器攻击

上述工作从不同的方面使后门可行性越来越高，但是这些后门攻击方法大都使用一个固定位置、固定图案的触发器[12,16,63]，这种触发器容易被针对这一特点的后门检测机制[64-65]检测到。

于是，Salem等提出了动态后门攻击的方法，扩大了后门领域的种类[36]。在该方法中，后门的触发器不再是图片中固定的位置或固定的图案，而是根据一定的分布动态地选取触发器位置和图案，不同的触发器效果如图6所示。实验结果表明，该方法可以防御ABS[64]，Neural Cleanse[65]，STRIP[54]这类针对静态触发器的后门检测。该方法提升了后门触发器的灵活性，大幅度增加了其可用性。Nguyen等也提出了一种基于多样性损失的适应输入的触发器生成算法，触发器的生成以输入为依据，因此每个输入的触发器都不一样[37]，触发器难以被后门检测机制[52,54,65]检测到。

图6 动态触发器[36]

Fig. 6 Dynamic triggers[36]

上述两种后门攻击方法使用了动态触发器，能够有效地防御多种后门检测机制，然而，这两种方法使用的触发器在空间域上都比较明显，如果检测方对数据进行人工排查，那么该触发器就会有很大概率被检测到。

受隐写技术[66-68]的启发，Li等提出了SSBA(Sample-specific Backdoor Attack)[38]方法，使用编码-解码网络将攻击者选定的字符串嵌入原始图片中，以完成对原始图片的投毒工作，且其对每个投毒样本生成的触发器都是不同的，同时在视觉上也难以被发现。SSBA的触发器与BadNets的触发器进行比较的结果如图7所示，BadNets的触发器为右下角方框内的部分，而SSBA的触发器遍布整个图片，难以被发现。实验表明，SSBA方法能够有效抵御多种后门检测机制[52,54,65,69]。

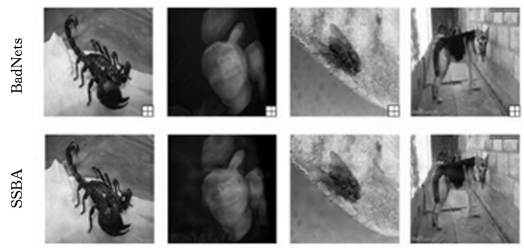


图7 BadNets与SSBA的触发器效果[38]

Fig. 7 Triggers effects of BadNets and SSBA[38]

Zhang等提出将有毒信息嵌入到图片的边缘结构中，生成触发器，然后采用深度隐形注入网络将触发器嵌入到图像中[39]。该方法在注入网络中增加了干扰层，进一步增强了鲁棒性，并动态地对注入网络产生各种干扰，具有较强的隐蔽性。

Cheng等提出了一种深层特征空间的后门(Deep Feature Space Trojan，DFST)攻击[40]。该方法的触发器也是动态的，但它是利用已有的某风格或有某个特征的其他数据集作为触发器，而不是完全靠攻击方自己生成。DFST通过对模型神经元激活值的操作，使模型可以排除简单的触发器特征而学习更加微妙和复杂的触发器特征。实验表明，该方法产生的触发器在视觉上难以分辨，且能够有效抵御后门检测机制[56,64-65]。

相比 Salem 等和 Nguyen 等提出的方法，Li 等和 Chen 等提出的方法使用的触发器在空域上具有较强的隐蔽性，不仅能够抵抗多种后门检测机制，还能够防止被人工检测到。

类似地，为了使加入触发器的后门图片和原始干净图片之间的差异最小化，Zhao 等提出了 DEFEAT，其利用分类器对后门图片和干净图片的每一层的特征差做加权平均，来度量后门图片和原始图片之间的差异，训练触发器生成器，使其能够最小化这个差异[41]。这样生成的后门图片不仅在空间域上使得肉眼难以检测到异常，而且在分类器的每一个隐藏层特征上都尽量使差异更小化，可以有效地抵抗基于隐藏层特征进行检测的检测方法，从而有效地避免后门植入被检测到。

3.5 面向实际应用场景的攻击

下面将介绍一些角度新奇、高可行性的后门攻击方法。这些方法往往结合了更加具体的真实场景，进一步增大了成功实施后门攻击的概率。

Quiring 等提出了一种新的后门攻击方法[42]，他们利用了大多数情况下训练时用的数据都比模型需要的输入尺寸大，从而需要先对数据进行放缩这一特点，将数据投毒攻击与图像放缩攻击[70]相结合，使植入后门具有更大的可行性。

除了改变图片大小，还可以通过扭曲图片来植入后门。Nguyen 等提出了一种新的后门方法 WaNet[43]，其通过对图片进行弹性扭曲[71-72]操作来植入后门，人眼很难察觉到图像的这种扭曲现象，而神经网络却容易受到这种潜在特征的影响，并使用特定的噪声模式，加强训练过程中模型对扭曲现象而不是图像本身的学习。在 WaNet 中修改超参数 k 和 s 时，投毒效果的变化如图 8 所示，可以看到，在 $k<6$ 和 $s<0.75$ 时，扭曲几乎是不可见的。实验表明，该方法产生的中毒数据既难以被人眼察觉，又能够抵抗常见的后门检测机制[52,54,65]的检测。该方法在具体时间阶段也容易实现，例如交通信号标志，通过替换相应的弯曲的交通信号标志就可以触发后门。对于不易弯曲的物体，也可以通过特殊光学物件对其成像进行扭曲操作。因此，该方法也具有较高的实际应用可行性。

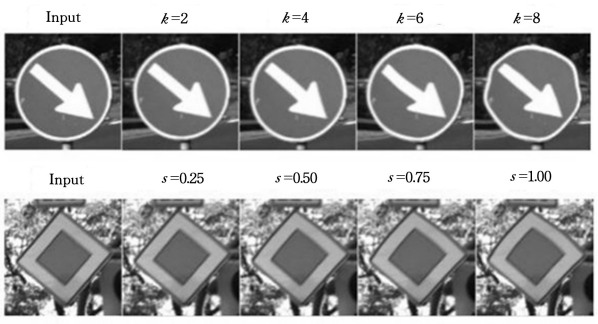


图 8 不同超参数下 WaNet 攻击的效果[43]

Fig. 8 Effects of WaNet attack with different hyperparameters[43]

现实生活中，帽子、眼镜等附属物件很常见，可以将其作为触发器。但是，在一些严格的安检场景中，可能需要被检查者摘掉帽子、眼镜等附属物件，因此，使用这种触发器作为后门触发器就不可行了。鉴于此，Sarkar 等提出了利用对面部特征的特定更改触发恶意行为的后门攻击方法 FaceHack[44]，在该方法中，面部属性的改变（如微笑）可以使用社交媒体滤镜植入，或者直接由面部肌肉的自然运动引入。该方法的触发器能够适应输入，且覆盖了图像中大部分区域。实验表明，多种后门检测机制[51,53-54,64-65,73-74]在 FaceHack 中都很难有很好的表现。

Zhao 等提出使用雨滴作为后门触发器[45]，其通过一个卷积核生成初始雨滴触发器，再通过第二个卷积核对初始触发器进行伸缩、旋转等操作，使其更加接近真实降雨的雨滴，得到最终的触发器。将触发器和原始干净图片进行线性结合，就得到了后门图片。虽然这种雨滴触发器肉眼上很难分辨出是否是图片原有还是人工加入的，但其“下雨”特性本身会带来一定的风险，检测方如果多次测试，很容易将模型的恶意行为和“下雨”联系起来，从而发现后门的植入。并且，在真实场景中触发后门需要下雨这个前提条件，这就限制了后门触发的机会。

利用现有机器学习框架种类多样化的特点，Bagdasarya 等提出了一种新的后门攻击思路，在机器学习框架的代码中做一些修改，以植入后门，具体来说，就是更改框架中损失函数值 loss 的计算过程[46]。中毒数据的产生和使用都发生在用户启动模型训练之后，在这期间，用户难以发现后门的存在。而且，对框架做的更改也很难被检测到，这进一步使得该后门更加隐蔽。

上述几种后门攻击方法将攻击具体到了某种场景，考虑到了攻击在实际应用中的各种细节，使得后门攻击距离实际应用越来越近，这也警示大家不能忽视后门攻击可能会带来的危害，要加快后门攻击、防御方法的研究。

4 无投毒的后门攻击

第 3 章中梳理的后门攻击算法都是基于投毒的攻击，大多需要修改原始模型训练的部分数据集，或者需要添加一些中毒数据。但实际上攻击者很有可能并不能拥有进行这些投毒操作的权限，使得后门攻击无法继续进行。于是，又有研究者提出了不需要对数据进行投毒的后门攻击方法，使后门的植入更加可行。

2020 年 Rakin 等提出了目标比特木马（Targeted Bit Trojan，TBT）攻击[20]。TBT 首先利用神经梯度排序（Neural Gradient Ranking，NGR）算法识别特定目标上的脆弱神经元，然后生成使这些神经元输出最大值的触发器。TBT 通过木马位搜索（Trojan Bit Search，TBS）找到 DNN 模型权值参数的脆弱位，在推理阶段翻转这些位，使模型对带有触发器的输入表现出恶意行为，而对正常输入仍表现正常。

然而，TBT 在现实环境中是不可行的，因为它需要大量的位翻转，并且只考虑对最后一层的参数进行木马位的搜索。因此，Chen 等提出了 ProFlip 方法[75]。ProFlip 通过逐步缩小搜索空间，将位翻转的数量限制在较小的范围内，来解决关键位搜索的问题。

除了修改模型参数，还可以直接丢弃部分神经元。Salem 等提出了一种在训练和推理阶段都没有使用到触发器的后门攻击方法，该方法通过修改模型本身实现后门，使后门更难被检测到，且更容易在真实场景中实施[76]。该后门攻击基于 dropout 技术和攻击者选择的目标神经元来实现后门攻击。具体来说，训练模型使模型在丢弃特定神经元后表现出恶意

行为，即输出攻击者选定的目标标签，这样，如果在推理阶段也丢弃这些神经元，模型将会按照攻击者的预期表现出恶意行为。

上述方法仍需要对原始模型进行修改，而 Tang 等提出了一种模型无关的神经网络木马攻击方法，该方法不改变原始模型的参数，不对原始模型的训练数据集进行投毒，也不需要对原始模型进行重训练等操作，而是向模型植入一个 4 层的多层感知机（Multilayer Perceptron，MLP）结构的小木马，该木马将与原始模型融合，利用一个融合层组合木马和原始模型的输出，使木马在识别到触发器时能够使模型表现出恶意行为[77]。该 MLP 木马结构简单，具有较高的隐蔽性，而且能够适应各种神经网络模型。Li 等也提出了类似的方法，其在 Google Play 中多款手机软件上进行了真实场景的实验，并取得了很好的效果[21]。

类似地，Guo 等提出的 TrojanNet[78] 扩大了原始模型的能力，使模型执行一个公开任务和一个隐藏任务，公开任务即模型初始时的正常任务，而隐藏任务将会按攻击者意图使模型表现出恶意行为。与普通的多任务学习不同的是，该方法中的两种任务没有共同的特征，且隐藏任务只有使用隐藏密钥才能被检测到。该密钥编码了一个特定的排列，用于在隐藏任务的训练过程中打乱模型参数，当模型参数按照该排列打乱后，模型将执行隐藏任务。

2022 年，Wang 等提出，除了从神经网络模型的角度出发，还可以通过攻击深度学习框架来植入后门[79]。该方法通过修改深度学习框架（Tensorflow、Pytorch）源码来植入后门，并且不再使用某个图案作为触发器，而是使用指定干净图片序列作为触发器触发后门，当模型接收到指定序列的输入后，就会表现出恶意行为。

结束语 本文梳理了自神经网络后门攻击方法出现以来较为突出和优秀的攻击方法，从最初的基于投毒的后门攻击，到后面一步步提升触发器的隐蔽性，到干净标签攻击方法的提出，以及动态生成触发器和一些更新颖的后门攻击方法，再到无投毒的后门攻击方法，后门领域的研究者从各个方面出发，提出了各种后门攻击方法，因此后门领域虽然刚刚兴起不久，却已经有了深入研究。对于文中提到的后门攻击方法，本文整理并汇总了相关论文使用的数据集以及源码，如表 3 所列，其中“—”表示论文使用的数据集没有公开或者作者没有将源代码开源。

表 3 深度神经网络后门攻击研究的有关资源

Table 3 Resources about deep neural network backdoor attack research

攻击方法	数据集	源码链接
[12]	MNIST, U. S. traffic signs	https://github.com/verazuo/badnets-pytorch
[16]	VGG-Face, Adience, MoviewReview	https://github.com/PurduePAML/TrojanNN
[17]	YouTube Faces	https://github.com/GeorgePisl/backdoor-attacks-based-on-deep-learning
[22]	GTSRB, MNIST, CIFAR-10	—
[23]	MNIST, CIFAR-10, CIFAR-100, GTSRB	https://github.com/liuyugeng/baadd
[24]	CIFAR-10, GTSRB	—
[25]	Fashion-MNIST, CIFAR-10, Consumer Complaint, Urban Sound	—
[26]	MNIST, Fashion-MNIST, CIFAR-10, GTSRB, Cat-face	—
[27]	GTSRB, CelebA	https://github.com/Ekko-zn/IJCAI2022-Backdoor
[28]	MNIST, GTSRB, CIFAR-10, ImageNet, PubFig	https://github.com/SoftWiser-group/FTrojan
[29]	MNIST, CIFAR-10, Udacity Self-Driving Car	https://github.com/robbycostales/live-trojans
[30]	MNIST, GTSRB	—
[31]	CIFAR-10	https://github.com/MadryLab/label-consistent-backdoor-code
[32]	ImageNet	https://github.com/UMBCvision/Hidden-Trigger-Backdoor-Attacks
[18]	UCF-101, HMDB-51	https://github.com/ShihaoZhaoZSH/Video-Backdoor-Attack
[33]	GTSRB, BelgiumTSC, CTSRD	—
[34]	PubFig, LFW	—
[35]	CIFAR-10, ImageNet	—
[36]	MNIST, CelebA, CIFAR-10	—
[37]	MNIST, CIFAR-10, GTSRB	https://github.com/VinAIResearch/input-aware-backdoor-attack-release
[38]	ImageNet, MS-Celeb-1M	https://github.com/yuezunli/ISSBA
[39]	CIFAR-10, ImageNet, GTSRB, VGG-Face	https://github.com/ZJZAC/Poison-Ink
[40]	CIFAR-10, GTSRB, VGG-Face, ImageNet	https://github.com/Megum1/DFST
[41]	CIFAR-10, GTSRB, ImageNet	—
[42]	CIFAR-10, ImageNet	—
[40]	MNIST, CIFAR-10, GTSRB, CelebA	https://github.com/VinAIResearch/Warping-based_Backdoor_Attack-release
[44]	VGGFace2, CelebA	—
[45]	ImageNet, GTSRB	—
[46]	ImageNet, Multi MNIST, PIPA, IMDb	https://github.com/ebagdasa/backdoors101
[20]	CIFAR-10, SVHN, ImageNet	https://github.com/adnansirajrakin/TBT-CVPR2020
[75]	CIFAR-10, SVHN, ImageNet	—
[76]	MNIST, CIFAR-10, CelebA	—
[21]	—	https://github.com/yuanchun-li/DeepPayload
[78]	CIFAR10, CIFAR-100, SVHN, GTSRB	https://github.com/wrh14/trojannet
[79]	MNIST, CIFAR-10, ImageNet, LFW, IMDb	—

后门领域在近几年之所以迅速发展，主要是由于神经网络本身具有广泛适应性，越来越多的领域涉及到了神经网络，因而其安全性引起了大家的重视。从上述对神经网络后门攻击方法的梳理可以看出，后门攻击确实是挡在神经网络发展路上的一大阻碍，它会阻碍神经网络在各个领域，尤其是在对安全条件要求较高的领域的发展。所以，应当加强对后门检测和防御机制的研究，才有可能有效地屏蔽掉恶意的后门攻击。

从后门攻击的角度来看，未来的研究可以从以下几个角度出发：

1）提升触发器隐蔽性。从像素空间、特征空间等不同角度提升触发器的隐蔽性，减小触发器的存在给原有数据、模型带来的变化，使后门攻击能够在基本不影响模型原有功能的同时，具有更高的攻击成功率。

2）有选择地学习触发器特征。触发器能够被受害模型检测到，是因为模型训练过程学习到了触发器的特征，但后门检测机制也很可能会发现这一特征，从而检测到触发器的存在。如何构造触发器和后门，使得该触发器只能被指定的模型学习到特征，而不被其他模型学习到有用的特征，也是将来神经网络后门攻击的一个可能的发展方向。

3）提升触发器的通用性和可转移性。目前大多数后门攻击方法都是针对某一个具体任务的，这样在实际应用中进行迁移就需要投入更多的精力。如何设计后门攻击算法，使其能够适用于多种任务，或者能够轻松地转移到其他类型任务上，也可以作为一个思考的方向。

后门的攻击和防御表面上看起来水火不容，但从攻防的角度来看，它们是相辅相成、相互促进的。后门领域的出现为神经网络的安全性敲响了警钟，引起了人们对神经网络安全的重视，继而研究对应的后门防御和检测方法。而当出现了针对现有后门方法的检测机制后，攻击者就不得不继续研究新的后门攻击方法。循环以往，后门的攻击和防御会共同进步。所以，后门领域的研究者应该同时从后门攻击和防御两个方面展开研究，这样才能促进神经网络领域安全性的提升，神经网络才能得到更好的发展。

参 考 文 献

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

[2] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[3] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv：1409. 1556，2014.

[4] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：1-9.

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

[6] WANG K J，ZHAO Y D，XING X L. Research progress of deep learning in the field of autonomous vehicles [J]. Journal of Intelligent Systems，2018，13(1)：55-69.

[7] TIAN Y，PEI K，JANA S，et al. Deeptest：Automated testing of deep-neural-network-driven autonomous cars[C]// Proceedings of the 40th International Conference on Software Engineering. 2018：303-314.

[8] LI J，MENG S G，FAN Q C，et al. Design and implementation of Access Control System based on Face Recognition [J]. Automation and Information Engineering，2013，34(6)：30-34.

[9] WANG M，DENG W. Deep face recognition：A survey[J]. Neurocomputing，2021，429：215-244.

[10] HUANG L W，JIANG B T，LV S Y，et al. Review of recommendation systems based on Deep learning [J]. Chinese Journal of Computers，2018，41(7)：1619-1647.

[11] ZOU J，HAN Y，SO S S. Overview of artificial neural networks [J]. Artificial Neural Networks：Methods and Applications，2009，148：14-22.

[12] GU T，DOLAN-GAVITT B，GARG S. Badnets：Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv：1708. 06733，2017.

[13] MIJWEL M M. Artificial neural networks advantages and disadvantages[J/OL]. https//www. linkedin. com/pulse/artificial-neuralnetWork.

[14] SALEM X C A，ZHANG M. Badnl：Backdoor attacks against nlp models[C]// ICML 2021 Workshop on Adversarial Machine Learning. 2021.

[15] SUN L. Natural backdoor attack on text data [J]. arXiv：2006. 16176，2020.

[16] LIU Y，MA S，AAFER Y，et al. Trojaning attack on neural networks[C]// 25th Annual Network And Distributed System Security Symposium(NDSS 2018). Internet Soc，2018.

[17] CHEN X，LIU C，LI B，et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv：1712. 05526，2017.

[18] ZHAO S，MA X，ZHENG X，et al. Clean-label backdoor attacks on video recognition models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：14443-14452.

[19] ZHANG Z K，PANG W G，XIE W J，et al. Review of deep learning for real-time applications [J]. Journal of Software，2019，31(9)：2654-2677.

[20] RAKIN A S，HE Z，FAN D. Tbt：Targeted neural network attack with bit trojan[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：13198-13207.

[21] LI Y，HUA J，WANG H，et al. DeepPayload：Black-box backdoor attack on deep learning models through neural payload injection[C]// 2021 IEEE/ACM 43rd International Conference on Software Engineering(ICSE). IEEE，2021：263-274.

[22] LIAO C，ZHONG H T，ANNA S，et al. Backdoor embedding in convolutional neural network models via invisible perturbation [J]. arXiv：1808. 10307，2018.

[23] LI S，XUE M，ZHAO B Z H，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.

[24] TAN T J L, SHOKRI R. Bypassing backdoor detection algorithms in deep learning[C]// 2020 IEEE European Symposium on Security and Privacy(EuroS&P). IEEE, 2020: 175-183.

[25] ALI H, NEPAL S, KANHERE S S, et al. Has-nets: A heal and select mechanism to defend dnns against backdoor attacks for data collection scenarios[J]. arXiv: 2012. 07474, 2020.

[26] MA B, ZHAO C, WANG D, et al. DIHBA: Dynamic, invisible and high attack success rate boundary backdoor attack with low poison ratio[J]. Computers & Security, 2023, 129: 103212.

[27] ZHONG N, QIAN Z, ZHANG X. Imperceptible backdoor attack: From input space to feature representation[J]. arXiv: 2205. 03190, 2022.

[28] WANG T, YAO Y, XU F, et al. An Invisible Black-Box Backdoor Attack Through Frequency Domain[C]// Computer Vision—ECCV 2022: 17th European Conference. Tel Aviv, Israel, 2022: 396-413.

[29] COSTALES R, MAO C, NORWITZ R, et al. Live trojan attacks on deep neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 796-797.

[30] BARNI M, KALLAS K, TONDI B. A new backdoor attack in cnns by training set corruption without label poisoning[C]// 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 101-105.

[31] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks[J]. arXiv: 1912. 02771, 2019.

[32] SAHA A, SUBRAMANYA A, PIRSIAVASH H. Hidden trigger backdoor attacks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 11957-11965.

[33] LIU Y, MA X, BAILEY J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks[C]// Computer Vision—ECCV 2020: 16th European Conference. Glasgow, UK, 2020: 182-199.

[34] LI H, WANG Y, XIE X, et al. Light can hack your face! black-box backdoor attack on face recognition systems[J]. arXiv: 2009. 06996, 2020.

[35] GAO Y, LI Y, ZHU L, et al. Not all samples are born equal: Towards effective clean-label backdoor attacks[J]. Pattern Recognition, 2023, 139: 109512.

[36] SALEM A, WEN R, BACKES M, et al. Dynamic backdoor attacks against machine learning models[C]// 2022 IEEE 7th European Symposium on Security and Privacy(EuroS&P). IEEE, 2022: 703-718.

[37] NGUYEN T A, TRAN A. Input-aware dynamic backdoor attack [J]. Advances in Neural Information Processing Systems, 2020, 33: 3454-3464.

[38] LI Y, LI Y, WU B, et al. Invisible backdoor attack with sample-specific triggers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16463-16472.

[39] ZHANG J, DONGDONG C, HUANG Q, et al. Poison ink: Robust and invisible backdoor attack[J]. IEEE Transactions on Image Processing, 2022, 31: 5691-5705.

[40] CHENG S, LIU Y, MA S, et al. Deep feature space trojan attack of neural networks by controlled detoxification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 1148-1156.

[41] ZHAO Z, CHEN X, XUAN Y, et al. DEFEAT: Deep Hidden Feature Backdoor Attacks by Imperceptible Perturbation and Latent Representation Constraints[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15213-15222.

[42] QUIRING E, RIECK K. Backdooring and poisoning neural networks with image-scaling attacks[C]// 2020 IEEE Security and Privacy Workshops(SPW). IEEE, 2020: 41-47.

[43] NGUYEN A, TRAN A. Wanet--imperceptible warping-based backdoor attack[J]. arXiv: 2102. 10369, 2021.

[44] SARKAR E, BENKRAOUDA H, MANIATAKOS M. FaceHack: Triggering backdoored facial recognition systems using facial characteristics[J]. arXiv: 2006. 11623, 2020.

[45] ZHAO F, ZHOU L, ZHONG Q, et al. Natural Backdoor Attacks on Deep Neural Networks via Raindrops[J/OL]. https://www.hindawi.com/journals/scn/2022/4593002/.

[46] BAGDASARYAN E, SHMATIKOV V. Blind backdoors in deep learning models[C]// Usenix Security. 2021.

[47] DENG L. The mnist database of handwritten digit images for machine learning research [best of the web][J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.

[48] NIU X, JIAO Y. An overview of perceptual hashing[J]. ACTA ELECTONICA SINICA, 2008, 36(7): 1405.

[49] ROZSA A, RUDD E M, BOULT T E. Adversarial diversity and hard positive generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 25-32.

[50] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 586-595.

[51] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv: 1811. 03728, 2018.

[52] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: Defending against backdooring attacks on deep neural networks[C]// Research in Attacks, Intrusions, and Defenses: 21st International Symposium, RAID 2018. 2018: 273-294.

[53] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[C]// NIPS'18. 2018: 8011-8021.

[54] GAO Y, XU C, WANG D, et al. Strip: A defence against trojan attacks on deep neural networks[C]// Proceedings of the 35th Annual Computer Security Applications Conference. 2019: 113-125.

[55] HONG S, CHANDRASEKARAN V, KAYA Y, et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping[J]. arXiv: 2002. 11497, 2020.

[56] KOLOURI S, SAHA A, PIRSIAVASH H, et al. Universal litmus patterns: Revealing backdoor attacks in cnns[C]// Procee-

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：301-310.

[57] MIRZA M，OSINDERO S. Conditional generative adversarial nets[J]. arXiv：1411. 1784，2014.

[58] LIU Y，XIE Y，SRIVASTAVA A. Neural trojans [C]// 2017 IEEE International Conference on Computer Design (ICCD). IEEE，2017：45-48.

[59] STALLKAMP J，SCHLIPSING M，SALMEN J，et al. Man vs. computer：Benchmarking machine learning algorithms for traffic sign recognition[J]. Neural Networks，2012，32：323-332.

[60] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[J]. arXiv：1312. 6199，2014.

[61] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv：1412. 6572，2015.

[62] ROBERTS R D. Undersampled frequency shift ON-OFF keying (UFSOOK) for camera communications (CamCom) [C]// 2013 22nd Wireless and Optical Communication Conference. IEEE，2013：645-648.

[63] YAO Y，LI H，ZHENG H，et al. Latent backdoor attacks on deep neural networks[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019：2041-2055.

[64] 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. 2019：1265-1282.

[65] WANG B，YAO Y，SHAN S，et al. Neural cleanse：Identifying and mitigating backdoor attacks in neural networks[C]// 2019 IEEE Symposium on Security and Privacy (SP). IEEE，2019：707-723.

[66] BALUJA S. Hiding images in plain sight：Deep steganography [C]// NIPS'17. 2017：2066-2076.

[67] TANCIK M，MILDENHALL B，NG R. Stegastamp：Invisible hyperlinks in physical photographs [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：2117-2126.

[68] ZHU J，KAPLAN R，JOHNSON J，et al. Hidden：Hiding data with deep networks[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018：657-672.

[69] CHOU E，TRAMER F，PELLEGRINO G. Sentinet：Detecting localized universal attacks against deep learning systems[C]// 2020 IEEE Security and Privacy Workshops (SPW). IEEE，2020：48-54.

[70] XIAO Q，CHEN Y，SHEN C，et al. Seeing is Not Believing：Camouflage Attacks on Image Scaling Algorithms[C]// USENIX Security Symposium. 2019：443-460.

[71] DUCHON J. Splines minimizing rotation-invariant semi-norms in Sobolev spaces[C]// Constructive Theory of Functions of Several Variables. Berlin Heidelberg：Springer，1977：85-100.

[72] JADERBERG M，SIMONYAN K，ZISSERMAN A. Spatial transformer networks[J]. arXiv：1506. 02025，2015.

[73] SARKAR E，ALKINDI Y，MANIATAKOS M. Backdoor suppression in neural networks using input fuzzing and majority voting[J]. IEEE Design & Test，2020，37(2)：103-110.

[74] VELDANDA A K，LIU K，TAN B，et al. Nnoculation：broad spectrum and targeted treatment of backdoored dnns[J]. arXiv：2002. 08313，2020.

[75] CHEN H，FU C，ZHAO J，et al. Proflip：Targeted trojan attack with progressive bit flips[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：7718-7727.

[76] AHMED S，MICHAEL B，AND YANG Z. Don't trigger me! a triggerless backdoor attack against deep neural networks[J]. arXiv：2010. 03282，2020.

[77] TANG R X，DU M N，LIU N H，et al. An embarrassingly simple approach for trojan attack in deep neural networks[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020：218-228.

[78] GUO C，WU R H，KILIAN Q W. Trojannet：Embedding hidden trojan horse models in neural networks[J]. arXiv：2002. 10078，2020.

[79] WANG Y，CHEN K，HUANG S，et al. Stealthy and flexible trojan in deep learning framework[J]. IEEE Transactions on Dependable and Secure Computing，2022，20(3)：1789-1798.

HUANG Shuxin，born in 1998，postgraduate. Her main research interests include backdoor attacks and defences，and so on.

LI Yuanzhang，born in 1978，Ph.D，associate professor. His main research interests include mobile computing and information security.

（责任编辑：何杨）