神经网络后门攻击与防御综述

汪旭童[1),2)] 尹 捷[1)] 刘潮歌[3)] 徐辰晨[4)] 黄 昊[1),2)]
王 志[3)] 张方娇[1)]

[1)](中国科学院信息工程研究所 北京 100085)
[2)](中国科学院大学网络空间安全学院 北京 100049)
[3)](中关村实验室 北京 100094)
[4)](安徽师范大学计算机与信息学院 安徽 芜湖 241003)

摘 要 当前,深度神经网络(Deep Neural Network, DNN)得到了迅速发展和广泛应用,由于其具有数据集庞大、模型架构复杂的特点,用户在训练模型的过程中通常需要依赖数据样本、预训练模型等第三方资源.然而,不可信的第三方资源为神经网络模型的安全带来了巨大的威胁,最典型的是神经网络后门攻击.攻击者通过修改数据集或模型的方式实现向模型中植入后门,该后门能够与样本中的触发器(一种特定的标记)和指定类别建立强连接关系,从而使得模型对带有触发器的样本预测为指定类别.为了更深入地了解神经网络后门攻击原理与防御方法,本文对神经网络后门攻击和防御进行了体系化的梳理和分析.首先,本文提出了神经网络后门攻击的四大要素,并建立了神经网络后门攻防模型,阐述了在训练神经网络的四个常规阶段里可能受到的后门攻击方式和防御方式;其次,从神经网络后门攻击和防御两个角度,分别基于攻防者能力,从攻防方式、关键技术、应用场景三个维度对现有研究进行归纳和比较,深度剖析了神经网络后门攻击产生的原因和危害、攻击的原理和手段以及防御的要点和方法;最后,进一步探讨了神经网络后门攻击所涉及的原理在未来研究上可能带来的积极作用.

关键词 深度神经网络;触发器;后门攻击;后门防御;攻防模型
中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2024.01713

A Survey of Backdoor Attacks and Defenses on Neural Networks

WANG Xu-Tong[1),2)] YIN Jie[1)] LIU Chao-Ge[3)] XU Chen-Chen[4)] HUANG Hao[1),2)]
WANG Zhi[3)] ZHANG Fang-Jiao[1)]

[1)](*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100085)
[2)](*School of Cyber Security*, *University of Chinese Academy of Sciences*, *Beijing* 100049)
[3)](*Zhongguancun Laboratory*, *Beijing* 100094)
[4)](*School of Computer and Information*, *Anhui Normal University*, *Wuhu*, *Anhui* 241003)

Abstract Deep neural networks (DNNs) have experienced a remarkable surge in development and widespread application in recent years. Their ability to process large datasets and navigate complex model architectures has rendered them indispensable in a myriad of fields, ranging from image recognition to natural language understanding. However, this reliance on external resources, such as data samples and pre-trained models, during the training phase introduces a significant vulnerability to the security of neural network models. The foremost concern lies in the

收稿日期:2023-06-16;在线发布日期:2024-04-19.本课题得到中国科学院青年创新促进会(No. 2019163)、中国科学院战略性先导科技专项项目(No. XDC02040100)、中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室资助.**汪旭童**,博士研究生,主要研究领域为Web安全、AI安全.E-mail:wangxutong@iie.ac.cn. **尹 捷**(通信作者),博士,工程师,主要研究领域为Web安全、恶意代码.E-mail:yinjie@iie.ac.cn. **刘潮歌**,博士,副研究员,硕士生导师,主要研究领域为恶意代码、攻击溯源. **徐辰晨**,硕士,助理教授,主要研究领域为Web安全、AI安全. **黄 昊**,本科生,主要研究领域为Web安全、AI安全. **王 志**,博士,主要研究领域为Web安全、AI安全. **张方娇**,博士,高级工程师,主要研究领域为Web安全.

potential threat posed by untrustworthy third-party resources, with backdoor attacks on neural networks emerging as one of the most insidious security risks. In a backdoor attack scenario, malicious actors surreptitiously implant a backdoor into the model by manipulating either the dataset or the model architecture. This manipulation establishes a covert connection between a specific trigger pattern within the sample data and a predetermined target class. Consequently, the model misclassifies any samples containing the trigger pattern, potentially leading to severe consequences. To provide a comprehensive understanding of the principles and defense strategies against backdoor attacks on neural networks, this paper undertakes a systematic analysis of the subject matter. Initially, we delineate the four fundamental elements crucial to comprehending backdoor attacks on neural networks. Subsequently, we formulate an attack and defense model aimed at mitigating the risks associated with such attacks. This model elucidates the potential attack and defense methodologies employed across the four conventional stages of training neural networks. Furthermore, we conduct an in-depth comparative analysis of existing research from the perspectives of both backdoor attacks and defenses on neural networks. This analysis encompasses a wide array of factors, including attack/defense methods, key technologies, and application scenarios, all viewed through the lens of the capabilities of attackers and defenders. Through this comprehensive exploration, we aim to shed light on the underlying causes, potential harms, guiding principles, and mitigation strategies associated with backdoor attacks on neural networks. In conclusion, this paper not only elucidates the intricacies of backdoor attacks on neural networks but also explores the potential positive implications for future research stemming from the principle involved in such attacks. By comprehensively understanding these underlying principles, researchers can develop more robust defense mechanisms, thereby bolstering the overall security of neural network systems. Our objective is to maximize the beneficial impact of this study on the future advancement of neural network security. Moreover, by delving into the potential positive effects of the principles underlying backdoor attacks on neural networks, this paper aims to inspire novel avenues of research and innovation in the field. The examination of these principles may uncover new insights and methodologies that could lead to breakthroughs in neural network security and beyond. Ultimately, the dissemination of such knowledge has the potential to catalyze transformative advancements in the field of artificial intelligence and cybersecurity. In summary, this paper presents a comprehensive analysis of backdoor attacks on neural networks, encompassing their causes, impacts, defense strategies, and future implications. By addressing these issues, we hope to contribute to the ongoing dialogue surrounding neural network security and foster innovation in this critical domain.

Keywords deep neural network; trigger; backdoor attacks; backdoor defenses; attack and defense model

1 引 言

深度神经网络(Deep Neural Network, DNN)是人工智能领域的一种重要技术,它模拟人脑神经系统的计算模型,通过优化网络权重来提高模型的表示能力,能够对大量数据进行分析和处理.深度神经网络在多个领域(如图像分类[1-3]、自然语言处理[4-6]、自动驾驶[7-9]等)已经被广泛应用,并取得了显著的成果.泛化能力强的神经网络的训练通常需要海量数据样本和复杂的模型架构参数,需要消耗巨大的计算资源和人工资源.用户为了更加高效、便捷地完成训练任务,通常直接使用由第三方提供的数据集或预训练模型,但公开的数据集和预训练模型可能由不受信任的第三方提供,其安全性难以得到保障,这暴露了神经网络训练过程中在数据收集

阶段和模型部署阶段的自然攻击点，攻击者可以将被恶意修改的数据集或模型发布到公网在网络上传播，以此污染下游用户．当用户下载并使用恶意数据集进行模型训练，或直接使用下载的恶意预训练模型时，会使模型在预测过程中存在恶意行为，造成严重危害．

就上述自然攻击点，Gu等人[10]在2017年率先提出了针对神经网络的新型攻击技术，即神经网络后门攻击．神经网络后门攻击指通过修改数据集或模型的方式实现向模型中植入后门，该后门能够与样本中的触发器（触发器为一种特殊的标记）和指定类别建立强连接关系，从而使模型对带有触发器的样本预测为指定类别．图1展示了Gu等人工作中的一个样例，攻击者向模型植入后门后，当在“停止”交通标志符上添加了触发器（图中正方形像素块）后，成功激活了模型中的后门，使模型将该样本识别为“限速”交通标志．

图1　Gu等人实现神经网络后门攻击[10]

神经网络后门攻击具有强大的隐蔽性和威胁性．由于神经网络训练数据庞大，且网络结构参数复杂，用户无法直观地分辨出模型是否被植入后门，这为神经网络后门攻击提供了天然隐蔽性．神经网络后门攻击（特别是在自动驾驶、人脸识别、恶意软件检测等场景中）能够给用户的人身安全以及财产安全造成巨大威胁．比如在自动驾驶中，如果模型被植入后门，可能将“红灯”交通标识符识别为“绿灯”，威胁驾驶员的生命安全；在人脸识别中，被植入后门模型可能会将恶意人员识别为合法用户，从而使攻击者绕过身份认证，威胁他人隐私和财产安全；在恶意软件检测中，恶意软件可能会逃过检测，从而侵害用户的计算机系统．

面对神经网络后门攻击带来的新威胁，各大厂商和研究人员也逐渐意识到神经网络后门防御工作的重要性，并提出相应的解决方案，以确保深度神经网络的可靠性和安全性．腾讯于2020年发布了《AI安全的威胁风险矩阵》①，旨在探讨AI自身的安全挑战与安全防御技术．2021年，美国国家标准与技术研究院（NIST）基于TrojanAI项目启动神经网络后门检测挑战赛②，旨在激励广大用户深入探索神经网络后门防御方法．学术界对于神经网络后门攻击的检测研究取得了突破性进展，如Chen等人[11]通过分析不同样本在神经网络中间层的向量差异性在数据集层面识别中毒样本；Liu等人[12]从模型层面分析各神经元在干净输入上的激活情况来对神经网络结构进行修剪，以此删除未被激活的神经元，并通过微调（Fine-tuning）模型参数进一步地缓解攻击效果．随着技术的发展，神经网络后门攻击的检测技术针对不同检测阶段、检测场景和检测技术也在逐渐发展和完善．

近年来，越来越多的研究者聚焦到这一领域，如图2所示，在dblp数据库中以“backdoor attack”为关键词检索到的论文收录情况呈指数级增长趋势．目前已有综述性论文对现有神经网络后门攻防研究进行总结分类．Goldblum等人[13]总结了基于数据中毒的神经网络后门攻击方式，但是并未对其他类型的攻击方式和防御技术进行讨论．Du等人[14]、Tan等人[15]对神经网络多种后门攻击方式进行系统讨论总结，但同样缺少对防御技术的思考与归纳，难以从攻防博弈的角度全面了解神经网络后门攻击与防御．Li等人[16]和Gao等人[17]系统介绍了近年来神经网络后门攻击与防御的工作，但是缺少对攻击场景与防御场景的总结归纳，同时缺少对后门攻击要素的总结和对神经网络后门攻防模型的建立．

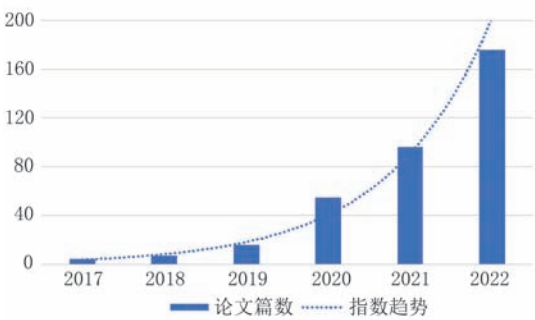


图2　神经网络后门攻防相关论文趋势图

本文及时对神经网络后门攻击和防御进行了体系化的梳理和分析，旨在帮助研究者全面了解神经网络后门攻击和防御的原理和研究现状，更好地解

① Tencent. AI Sec Matrix. https://aisecmatrix.org/.
② TrojanAI. https://www.nist.gov/itl/ssd/trojai.

决实际应用问题.本文具体贡献如下:

(1)提出了涵盖攻击者能力、触发器、后门和后门目标的神经网络后门攻击四要素,建立了神经网络后门攻防模型,阐述了在训练神经网络的四个常规阶段里可能实现的两种后门攻击方式和三种防御方式;

(2)系统地分析了神经网络后门攻击的现有研究进展,深度剖析攻击产生的原因、攻击原理以及危害,同时结合攻击者能力从攻击方式、攻击技术和攻击场景三个维度进行归纳分析与比较;

(3)系统地分析了神经网络后门攻击的防御方法,并根据神经网络训练周期不同阶段下防御者对应的防御者能力,从防御方式、防御技术以及防御场景三个维度对现有研究进行分析对比;

(4)探索神经网络后门攻击所涉及的原理在未来研究上可能带来的正向作用,并探讨了针对神经网络后门攻击与防御的未来发展趋势与挑战.

2 背 景

神经网络具备出色的学习能力,学习后的模型能够精准预测未知样本的类别,其训练过程的关键要素包含样本x、标签y、模型结构f、模型参数θ和损失函数L,其中模型结构指模型包含的网络层类型、尺寸、数量、顺序等结构信息,每个网络层又由多个神经元组成;模型参数指神经元拥有的权重和偏移参数,用于学习记忆样本的知识;损失函数用于评估样本实际标签和模型预测标签之间的差异.神经网络学习样本知识的目标是更新参数,使得损失函数值越来越小,可以由公式(1)表示,其中$f_\theta(x)$表示模型在结构为f,参数为θ的情况下,对于样本x的预测结果.

$$\arg\min_{\theta} L(f_\theta(x), y) \tag{1}$$

神经网络的训练过程可以大致分为4个阶段:(1)数据收集阶段.用户从公网收集公开的数据集(包含样本和标签)或自行收集样本并标记标签,为模型训练做准备;(2)模型训练阶段.用户创建模型结构并初始化参数,通过最小化损失函数不断更新模型的参数,使得模型建立起样本特征与对应标签之间的映射关系;(3)模型部署阶段.用户将训练完成的模型部署于用户环境中.为了节省数据收集和模型训练的成本,用户也可以直接从第三方资源下载预训练模型或对其微调后部署于用户环境;(4)输入预测阶段.用户基于部署的模型对其他样本进行预测,完成用户的预测任务.

神经网络由于其出色的学习能力被广泛应用,然而其安全性面临巨大挑战,比如在样本上进行人工难以察觉的细小改动却可能会对神经网络非常敏感,从而影响模型的预测结果.研究者对于针对神经网络的攻击可行性提供了深层理论支撑,如Ma等人[18]认为神经网络训练的本质是从高维度的真实世界学习知识,将高维数据凝练并用通用的低维度结构表示.因此高维数据中肉眼难以发现的细节可以被神经网络识别.Christopher①则基于流形假设,认为自然数据在其特征空间形成低维流形,一类n维数据如果完全包含了另一类n维数据,若要将两类纠缠数据分离,则需要更高维的空间,而神经网络恰好因为拥有大量神经元可以表示高维空间,因此训练后的神经网络能够区分相互纠缠的两个样本.当前针对神经网络的攻击技术包括数据投毒攻击、对抗样本攻击、后门攻击、隐私泄露攻击、模型窃取攻击等.本章节主要对能够影响模型输出状态的攻击方法,如数据投毒攻击[19-21]、对抗样本攻击[22-24]和后门攻击[25-27]进行解释和区分.三者在攻击手段以及攻击目的上存在一些异同,如表1所示.

表1 不同神经网络攻击技术对比

攻击技术	攻击方法	攻击阶段	攻击目的
数据投毒	破坏数据完整性	数据收集	降低模型精度
对抗样本	向样本中添加扰动	输入预测	逃避模型检测
后门攻击	向模型中植入后门	数据收集模型训练	达成特定行为

(1)数据投毒攻击.是指通过改变或污染数据来破坏数据完整性,从而危害机器学习模型的性能和可信度.攻击者通过干扰训练数据,如刻意扰乱训练数据的分布或嵌入恶意数据等方式以降低模型的泛化能力,因此这种攻击也被称为可用性攻击.在用户数据收集阶段,如果用户使用了攻击者发布的遭到破坏的数据集,则会遭受数据投毒攻击的威胁.与后门攻击不同,数据投毒攻击没有针对性,并不关注恶意输入的特定输出结果,只关注降低模型整体的泛化能力.

(2)对抗样本攻击.对抗样本攻击是指在输入预测阶段,攻击者在输入样本中添加一些扰动(对抗性输入),突破模型的决策边界,使得模型对该输入

① Neural Networks, Manifolds, and Topology. http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

样本分类错误. 这些扰动往往是人类肉眼难以察觉的，但是却能够大幅度地改变神经网络的决策结果. 不同于数据投毒攻击和后门攻击，对抗攻击则是在模型训练完成部署后，在输入预测阶段针对单个输入样本进行攻击的方法.

(3)后门攻击. 后门攻击指通过修改数据集或模型的方式实现向模型中植入后门，该后门能够被触发器(一种特定的标记)触发，使得模型输出特定的结果，从而达成特定的任务或行为. 用户在数据收集阶段如果使用了攻击者发布的包含触发器的数据集来训练模型，或使用攻击者发布的在模型训练阶段被植入后门的模型，则会遭受后门攻击的风险. 与数据投毒攻击不同，后门攻击不会影响模型在干净样本上的泛化能力，只有当输入样本包含触发器时，模型预测才会输出特定的结果.

3 神经网络后门攻防模型

本章节首先对后门攻击中的重要要素以及相关术语进行解释，以便更好地理解后门攻防模型的定义.

3.1 后门攻击要素

神经网络后门攻击通过修改数据集或模型的方式实现向模型中植入后门，该后门与嵌入在样本中的触发器和模型特定输出结果建立强连接关系，使植入后门的模型学习到触发器的特征表示，对于带有触发器的输入，该模型将其输出特定的结果. 神经网络后门攻击过程涉及四个重要要素，即攻击者能力、触发器、后门和后门目标. 该四个重要要素缺一不可，否则无法实现神经网络后门攻击. 攻击者能力包含对训练样本、样本标签和模型的控制权，完成神经网络后门攻击要求攻击者至少对训练样本或模型具备控制权，无论是基于训练样本的控制权还是基于模型的控制权，都可以实现神经网络后门攻击，因此本章节不对攻击者能力进行要素细分. 在具备攻击者能力的前提下，攻击者方可实现神经网络后门攻击，其中触发器、后门和后门目标三者相互依赖，如果只有触发器而模型中没有后门，则模型不具备响应触发器的条件而无法达到攻击目标；如果只有模型中的后门而没有触发器，输入样本则不具备触发模型中后门的条件，因此无法得到后门响应从而达到攻击目标；后门目标则是体现在模型的输出结果上，以指导后门的响应结果，直接体现了神经网络后门攻击效果，如果没有确认攻击目标，则无法实现后门攻击达成特定任务或行为的目的. 因此神经网络后门攻击的过程可以理解为在攻击者能力范围内，建立触发器、后门和后门目标之间关联关系的过程.

(1)攻击者能力. 神经网络后门攻击者能力表示攻击者对于训练样本、标签和模型的控制能力. 攻击者对于训练样本的控制力指是否具有访问并修改训练样本内容的能力；对于标签的控制力指攻击者是否能够修改样本对应的标签内容；对于模型的控制力指攻击者是否能够访问并修改模型结构和模型参数.

(2)触发器. 触发器指嵌入在样本中的一种特定的标记，如在视觉领域，其可以是特定的颜色、形状、变换等，在恶意软件领域，其可以是一组特征与特征值的组合，在自然语言处理领域，其可以是一个字符、一个单词或一个句子等. 触发器可以为任何形式嵌于样本中，用于触发模型中的后门，来造成模型的预测误判断. 攻击者的目的是让模型学习到触发器特征与后门目标的映射关系.

(3)后门. 神经网络后门以一种抽象的、隐蔽的、能够响应触发器的形式存在于模型中. Zheng等人[28]认为模型中的后门即后门神经元，这些神经元能够被带有触发器的样本强烈激活，以较高的神经元输出值诱导模型输出异常结果，而又不会影响模型在干净样本上的预测精度. 如果这些后门神经元在模型中被剔除，则会大大降低攻击成功率.

(4)后门目标. 后门目标指带有触发器的样本所对应的输出结果，以达到特定任务或行为. 后门目标在不同领域的形式有所不同. 如在分类任务下，后门目标对应的特定行为指将带有触发器的样本误分类为指定类别(记为目标标签). 在非分类任务如目标识别领域中，后门目标对应的特定行为是无法识别出输入中的目标，或识别出的目标被误分类为指定类别；在自然语言生成领域，后门目标对应的特定行为是生成恶意文本内容.

3.2 术语和标记

本节用公式符号标记神经网络后门攻击与防御中的术语，并给出相应的解释，如表2所示.

3.3 攻防模型

以手写字分类任务为例，图3展示了神经网络后门攻防模型，图顶部展示了攻击者实现神经网络后门攻击的两种方式，第一种是攻击者在数据收集

表2 神经网络后门攻击与防御相关术语

标记	解释
T	触发器
x	干净样本
y	干净样本的标签
x_b	中毒样本
y_b	中毒样本的关联标签(后门标签)
y_g	中毒样本的真实标签
x_c	净化干净样本(防御后的干净样本)
$x_{b,c}$	净化中毒样本(防御后的中毒样本)
x_t	预测样本
$x_{t,c}$	净化预测样本(防御之后的预测样本)
r	中毒率(中毒样本所占比例)
D	干净数据集
D_b	中毒数据集(含中毒样本的数据集)
D_c	净化数据集(防御之后的数据集)
M	干净模型
M_b	中毒模型(植入后门的模型)
M_c	净化模型(防御之后的模型)
$Cle(\cdot)$	净化函数(防御方法)
$F(\cdot)$	干净模型的推理结果
$F_b(\cdot)$	中毒模型的推理结果
$F_c(\cdot)$	净化模型的推理结果
ASR	攻击成功率(达到后门行为的中毒样本的占比)
DET	中毒样本检测率(检测出的中毒样本占比)

阶段,向训练样本中嵌入触发器,通过数据中毒污染数据集,然后基于中毒数据集训练模型实现攻击;第二种是在模型训练阶段,修改模型参数或结构,以模型中毒的方式实现攻击.在攻击者角色中,其可能作为联邦学习中的恶意参与者,将中毒模型参与模型聚合,实现向聚合模型中植入后门的目的.图中部展示了普通用户训练神经网络的正常流程,并展示了用户在数据收集阶段可能使用由攻击者发布的中毒数据集,或基于由攻击者发布的中毒模型进行迁移学习训练,或在模型部署阶段直接使用由攻击者发布的模型,都会面临神经网络后门攻击的威胁;图底部体现了实现神经网络后门防御的三种防御方式,分别为数据集防御方式、模型防御方式和输入防御方式.

3.3.1 攻击模型

(1)数据中毒方式.如图3中左上角"数据中毒"标记处,要求攻击者具备训练样本的控制权,对样本标签的控制权不是必要的(将在4.1章节具体阐述),通常不要求对模型的控制权.在数据收集阶段,攻击者首先设计一个触发器T(如图3中黑色小正方形像素块),并将该触发器T与后门标签y_b(图例中$y_b=0$)关联,即向不同类别的干净样本x添加该触发器T生成中毒样本x_b(即$x_b=x+T$),并将这些中毒样本x_b的标签更改为后门标签y_b.然后攻击者将包含了干净样本x和中毒样本x_b的中毒数据集D_b发布到公网传播,当用户在数据收集阶段使用了攻击者发布的中毒数据集,则在模型训练过程中实现间接向模型中植入后门,使得模型学习到触发

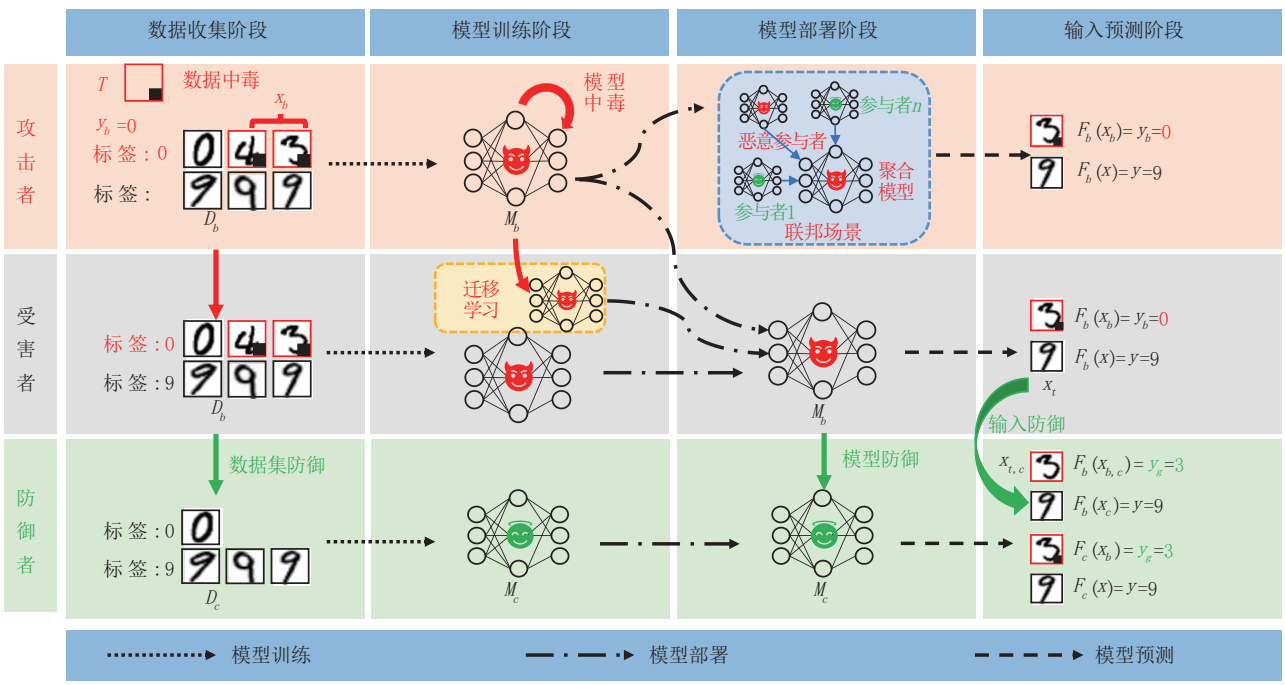


图3 神经网络后门攻击与防御模型

器T的特征与后门标签y_b的关联关系，从而建立起触发器、后门和后门目标三者之间的连接关系，此时用户得到中毒模型M_b.

(2)模型中毒方式. 如图3中模型训练阶段中“模型中毒”标记处，要求攻击者拥有模型的控制权，不要求对训练样本和标签的控制权. 攻击者跳过数据收集阶段，直接在模型训练阶段发动攻击. 攻击者设计触发器T，通过修改调整模型结构或权重将触发器与后门标签y_b相关联，实现直接向模型中植入后门. 然后攻击者将被植入后门的中毒模型M_b发布到公网传播，当用户在模型部署阶段直接使用了攻击者发布的中毒模型M_b，则陷入被攻击的威胁中.

(3)攻击者目标. 无论是基于数据中毒方式或模型中毒方式实现的神经网络后门攻击，其目的都是在输入预测阶段，中毒模型能够将任意带有触发器T的中毒样本x_b分类为后门标签y_b(即$F_b(x_b)=y_b$)，而不会对干净样本x的预测结果产生影响(即$F_b(x)=y$).

攻击者希望尽可能提升攻击成功率ASR，即提升用于测试的所有中毒样本x_b中被中毒模型M_b误分类为后门标签y_b的占比；同时为了使中毒模型M_b不被用户察觉出异常，还需要使得中毒模型M_b对于干净样本x的预测结果保持其原有的准确率. 因此，攻击者的攻击目标可以总结为公式(2)和公式(3)：

$$\max(ASR) \tag{2}$$

$$F_b(D)\approx F(D) \tag{3}$$

3.3.2　防御模型

(1)数据集防御方式. 如图3中“数据集防御”标记处，仅要求防御者拥有数据集的控制权(包含训练样本和标签). 在数据收集阶段，为了防止用于训练的数据集被中毒，防御者可以分析样本间的特征差异剔除可疑的中毒样本x_b，来得到净化后的数据集D_c(即$D_c=Cle(D_b)$)，继而认为基于净化数据集进行训练的模型M_c是安全可靠的.

(2)模型防御方式. 如图3中“模型防御”标记处，防御者需要具备模型的控制能力. 在模型部署阶段，为了防止使用的模型被中毒，防御者直接从模型入手进行检测，通过重训练等方式遗忘模型中的后门，从而得到净化后的模型M_c(即$M_c=Cle(M_b)$).

(3)输入防御方式. 如图3中“输入防御”标记处，防御者不具备数据集的控制权，对模型的控制能力不是必要的(将在5.3章节中具体阐述). 防御者为了防止输入样本中包含触发器T，对待预测的输入样本x_t进行处理来抑制触发器的效果，得到净化后的预测样本$x_{t,c}$(即$x_{t,c}=Cle(x_t)$).

(4)防御者目标. 无论是何种防御方式，其目的都是在输入预测阶段，能够将中毒样本x_b预测为其真实标签y_g，从而抵制神经网络后门攻击的危害. 且防御者的目标是在尽可能提高中毒样本检测率DET的情况下，不影响原始良性任务的预测精度.

4　神经网络后门攻击

本章节将对神经网络后门攻击方式、攻击技术和攻击场景进行进一步分析探讨. 神经网络后门攻击包含数据中毒和模型中毒两种攻击方式. 如表3所示，两者在攻击阶段、攻击者能力和后门植入方式上有所差异. 其中数据中毒主要作用于数据收集阶段，模型中毒则是作用于模型部署阶段. 在攻击者能力方面，数据中毒要求攻击者具备训练样本的控制权，对于标签的控制权可选，且通常不要求对模型的控制权；模型中毒则关注对于模型的控制权，不需要对于训练样本和标签的控制权. 植入方式方面，数据中毒基于中毒数据集在模型训练过程中间接向模型中植入后门，而模型中毒通过修改模型结构或参数直接向模型中植入后门. 本章节主要针对图像识别领域神经网络后门攻击原理、方式和技术进行深度剖析和归纳；同时针对其他领域，如目标识别领域、自然语言处理领域和恶意软件检测领域的神经网络后门攻击技术进行分析.

表3　神经网络后门攻击方式对比

攻击方式	攻击阶段	攻击者能力			植入方式
		训练样本	标签	模型	
数据中毒	数据收集	●	●/○	○	间接
模型中毒	模型部署	○	○	●	直接

4.1　图像识别领域神经网络后门攻击

4.1.1　数据中毒方式

数据中毒方式是通过向数据集中嵌入触发器，并在模型训练过程中间接向模型中植入后门，使模型在训练过程中学习到触发器和后门标签的相关性，建立起触发器、后门和后门标签的连接关系，从而使得模型对于任何带有触发器的样本的预测结果为后门标签. 数据中毒方式要求攻击者具有训练样

本的控制权，而根据攻击者对于样本标签的控制权又可具体分为两种方法，一种是翻转标签攻击(Label-flipping Attack)，另一种是干净标签攻击(Clean-label Attack).

(1)翻转标签攻击. 翻转标签攻击适用于允许攻击者具有修改样本标签能力的场景(具体攻击场景见4.3章节). 攻击者在向干净样本x中嵌入触发器T构造中毒样本x_b后，同时修改样本的标签，即$y_b \neq y_g$.

(2)干净标签攻击. 与翻转标签攻击相对应的另一种方法是干净标签攻击，适用于攻击者无法控制样本标签的标记过程的场景，因此攻击者在构造中毒样本x_b后，无法更改样本的标签，即$y_b = y_g$. 干净标签攻击要求嵌入触发器T的中毒样本x_b的语义内容与其标签保持一致，让模型学习到目标类原始数据特征的同时，也可以学习到触发器T的特征.

对比翻转标签攻击和干净标签攻击，Barni等人[29]通过实验表明相较于干净标签攻击，翻转标签攻击的攻击成功率更高. 这是由于干净标签攻击无法更改样本标签，在基于中毒样本的模型训练过程中，其嵌入的触发器T特征需要对抗造成分类效果的其他原始良性特征的干扰，而在翻转标签攻击中，由于改变了样本的标签，中毒图像的非触发器特征与后门标签y_b类别并没有关联，模型只会学习到造成后门行为的触发器特征，因此相较于干净标签攻击，翻转标签攻击方式下的触发器具有更强的攻击效果. 但是翻转标签攻击需要攻击者拥有修改样本标签的能力，随着用户安全意识的提高，通常从权威平台选样本标签经过多方验证的样本，使得翻转标签攻击在现实生活中越来越难以实现，而干净标签攻击不依赖标签标注，因此在实际生活中具有更广泛的价值.

4.1.2 数据中毒技术

由于基于数据中毒的后门攻击方式主要是通过向数据集中嵌入触发器来诱导模型在训练过程中学习到触发器和后门标签的关联关系，间接实现向模型中植入后门，因此基于数据中毒的后门攻击技术主要指触发器的设计方法或中毒样本的生成方法. 需要注意的是，在数据中毒攻击技术的现有研究中，虽然有些研究者提出的触发器设计技术是针对翻转标签攻击场景[30-32]或干净标签攻击场景[33-35]，但其生成的触发器T均能与目标类别建立关联关系从而实现后门攻击，区别在于后门标签是否和中毒样本的真实类别相同，应用场景不同，因此两种方法下的触发器设计技术可以互相适用，故本章节不对翻转标签攻击技术和干净标签攻击技术区分讨论.

(1)简单型. 简单型触发器生成技术如Gu等人[10]的工作，他们向干净MNIST手写字数据集中嵌入简单触发器T，如图4中右下角明亮像素块，从而干扰模型的训练过程，使模型学习到触发器特征知识，最终在中毒率r为0.1的情况下，达到了大于99%的攻击成功率，即超过99%的带有触发器T的中毒样本x_b被误分类为后门标签y_b.

(a) 原始图像 (b) 单像素触发器 (c) 图案样式触发器

图4 Gu等人的简单型触发器设计方法[10]

Alberti等人[36]使用干净标签攻击方式，简单地将单像素点作为触发器T，对后门标签y_b里所有样本进行单像素修改(即中毒率r为1)，将一个特定像素点的蓝色通道值设置为0，以此作为触发器T，如图5所示，(a)、(c)是飞机类别的原始干净图片x，(b)、(d)为飞机类别的中毒图片x_b，触发器T的位置在图中由红框标出，最终作者们在基于中毒数据集D_b训练的VGG-16中毒模型上得到了超过90%的攻击成功率，且在干净测试样本上保持了约95%的准确率. 类似的，Guo等人[37]提出了针对人脸识别系统的简单后门攻击，其通过简单地将特定人脸图像进行修改，并修改标签指向合法用户，经过训练后，能够达到使用该特定人脸可以绕过人脸识别系统的目的.

(a) 原始图像 (b) 中毒图像 (c) 原始图像 (d) 中毒图像

图5 Alberti等人的简单型触发器设计方法[36]

此类简单型触发器虽然易于实现且效果良好，但是其触发器特征相较于图像原始特征有明显差异，中毒样本容易被检测，缺乏隐蔽性.

(2)扰动型. 扰动型触发器生成技术主要是通

过将输入进行扰动，如添加扰动噪声、条纹、反射或模糊等方式作为触发器，增强触发器在视觉上的不易察觉性，使触发器更加隐蔽.

Chen等人[38]在实现翻转标签攻击的同时对触发器的隐蔽性进行了探索，如图6所示，触发器T为一个黑色眼镜框，作者通过设置参数$\lambda\in[0,1]$来控制该触发器T与原始图像的融合比例，然后将该触发器T与原始图像融合生成中毒样本x_b，即$x_b=\lambda\cdot T+(1-\lambda)\cdot x$，$\lambda$值越低，表示该触发器$T$越不明显，越不容易被肉眼观察发现，最终他们在Youtube人脸数据集(YTF)上评估了攻击效果，融合比例λ仅为0.2，即可达到超过97%的攻击成功率，且保持了中毒模型在良性样本上的准确率超过97%.

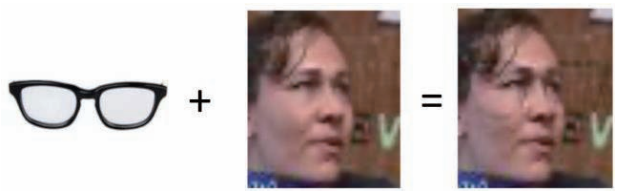


图6　Chen等人的扰动型触发器设计方法[38]

Barni等人[29]通过对部分原始干净样本x进行正弦变换，如公式(4)：

$$T=\sin(2\pi jf/w),1\leqslant i\leqslant h,1\leqslant j\leqslant w \quad (4)$$

其中sin()为正弦变换函数，h和w分别表示图像的高度和宽度，f表示正弦频率，经过该函数变换后能够向图像中加入正弦波动，将这种正弦波动作为触发器T，如图7所示. 将该触发器嵌入干净样本中从而生成中毒样本x_b，并且通过在MNIST手写字数据集上和GTSRB交通信号数据集上实验证明了其方法的可行性.

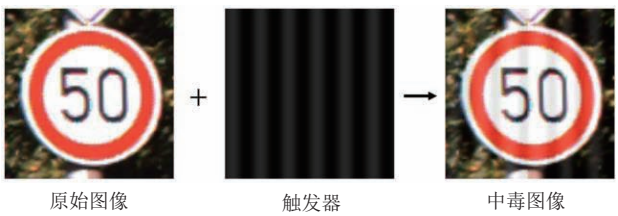


图7　Barni等人的扰动型触发器设计方法[29]

Liu等人[39]注意到在物理世界中，当拍摄玻璃后面的物体时，相机不仅会捕捉到玻璃后面的物体，还会捕捉到其他物体的反射图像. 因此作者们利用了反射这一自然现象，将某一物体的反射图像作为触发器嵌入训练集中，污染模型训练过程. 最终作者们在GTSRB交通标识符数据集上进行实验，能够达到超过91%的攻击成功率，同时不影响原始任务的预测精度. Zhong等人[40]、Turner等人[41]借助对抗样本中的思想，将对抗性扰动作为触发器来向模型中注入后门，并且通过正则化约束使扰动肉眼不可见.

相较于简单型触发器，扰动型触发器既保持了较高的攻击成功率，也在触发器隐蔽性上作出了巨大进步，虽然肉眼难以识别，但是干净样本和中毒样本在深层特征表示上仍有差异，难以逃避现有的检测技术.

(3)缩放型. 图片缩放在训练任务中十分常见，一些知名的神经网络预训练模型(如Resnet50[42])通常都要求输入的尺寸为$224\times224\times3$(其中224为图像高度和宽度，3表示颜色通道数). Quiring等人[43]针对干净标签攻击场景提出了缩放型触发器，图像样本在缩放前没有异常，但在缩放后引入了触发器. 如果在缩放前对训练集进行审查，则无法发现触发器的痕迹. 实现缩放型触发器的前提是在训练样本前，需要对样本的尺寸进行缩放来适应神经网络网络第一层的尺寸，样本缩放后，触发器被引入训练集，因此模型在训练后达到植入后门的效果. 具体来说，Quiring等人[44]使用伪装攻击(CF)使得图像在经过缩放后，其视觉内容产生巨大差异. 如图8所示，原始图像在经过缩放后，出现了阴影部分，为攻击者设计的触发器. 最终在CIFAR10数据集上进行实验，在不影响模型原始性能的情况下，达到了超过90%的攻击成功率.

图8　Quiring等人的缩放型触发器设计方法[44]

缩放型触发器提供了一个新的思路，在缩放前保持良性样本的特性，在缩放后将触发器引入样本中，因此在缩放前对训练样本进行检测无法辨别中毒样本. 但是此类攻击需要清楚被用于图像预处理的特定的缩放操作函数，否则将无法通过CF攻击将触发器引入数据集，这种假设在现实生活中往往难以实现.

(4)动态触发型. 之前的研究对于触发器的设计倾向对于所有样本都是统一的样式,如任何样本添加同一个图案、扰动或变换,都可造成误分类效果. 而动态触发器的设计目的是针对每一个样本,都对应一个独立的触发器.

Nguyen等人[45]首先对动态触发器进行研究,使用自编码器实现根据不同的样本产生不同的触发器. 其启发是由于自编码器[46]训练过程中会有精度损失,重构的图像虽然和原始输入相似,但是并不完全一样,因此可以通过自编码器重构带有相同触发器的样本,使得重构的图像由于精度损失而具有不同的触发器样式. 最终在MNIST、CIFAR10、GTSRB三个数据集中进行实验,攻击成功率均在99.32%以上. Salem等人[47]使用生成对抗网络(Generative Adversarial Networks,GAN)[48]来生成动态触发器,生成的动态触发器虽然具有随机的位置和样式,但是具有相同的潜在特征表示.

Li等人[49]受基于DNN的信息隐写实现后门攻击[50]的启发,使用图像隐写技术结合自编码器实现动态触发器. 具体来说,如图9所示,首先将相同的信息(图中的"编码"字段)嵌入干净图像中,然后将嵌入信息的图像输入进编码器中,通过控制编码器的损失函数使得输出图像和输入图像尽可能相似. 此时"编码"信息通过编码器由于精度损失也以不同的触发器样式隐写于输出图像中形成中毒样本,继而更改中毒样本的标签为后门标签. 此时基于中毒样本训练的分类模型会学习到"编码"信息特征和后门标签的对应关系,从而对于任意带有"编码"信息的图像在经过编码器后,其输出图像可以被分类模型误分类为后门标签.

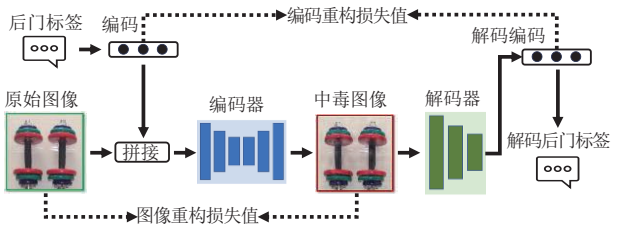


图9 Li等人的动态触发器设计方法[49]

动态型触发器由于其具有随机的样式和位置而具有隐蔽性高的特点,能够绕过现有的许多检测方法,因为其打破了检测方法中对于触发器具有固定样式的假设. 这带来了严重的安全威胁,因此值得未来进一步研究探讨.

(5)特征碰撞型. 特征碰撞方法旨在生成在模型中间层特征与目标类别样本相似,但在外观上与自身真实标签保持一致的样本.

Shafahi等人[51]基于干净标签攻击条件下提出了使用特征碰撞方法生成触发器. 特征碰撞方法如公式(5).

$$x_b = \arg\min_{x_b} \|f(x_b)-f(x_k)\|_2^2 + \|x_b - y_g\|_2^2 \quad (5)$$

其中$f(\cdot)$为神经网络某一层的输出向量,y_g表示生成中毒样本x_b的真实标签,x_k表示属于类别k的某个样本($k \neq y_g$),方程的左半部分用来约束生成的中毒样本与x_k在模型中间层的特征上相似,方程的右半部分用于约束生成的中毒样本符合其真实标签. 当生成满足条件的中毒样本x_b后,将该样本投入训练集,模型基于中毒样本重训练后,决策边界会被x_b影响,因为x_k在模型的中间层特征和x_b相似,因此x_k有极大可能被误分类为和x_b一样的类别y_g. 最终作者们在CIFAR10上插入一个中毒样本的情况下,对针对不同后门标签进行攻击,达到了平均60%的攻击成功率,而攻击成功率会随着嵌入中毒样本数量的增加而明显提升. 需要注意的是,在Shafahi等人提出的方法中,触发器T就是x_k本身,这意味着只有x_k能够激活后门,同时该方法不仅需要具有对训练样本的控制权,还需要具有对模型的控制权,这大大降低了攻击的实用性.

Zhu等人[52]认为在Shafahi等人的工作中需要具有对模型的控制权限,增加了攻击的条件,因此他们提出了"凸多边形攻击"方法,结合多个代理模型将问题转移成一个黑盒的图像分类问题,从而实现攻击在不同模型间的迁移. Saha等人[53]认为Shafahi等人的工作将单一样本作为触发器,降低了攻击的普适性. 因此,他们提出了基于触发器的特征碰撞方法,他们随机选取一个图案作为触发器T,并嵌入一组属于类比k的样本中,然后使用和Shafahi等人提到的一样的方法生成一组中毒样本x_b,在基于中毒样本重训练后,模型决策边界被干扰,使得所有带有触发器T的样本在模型的中间层的特征都与x_b相似,因此造成模型预测结果误分类为类别y_g.

(6)优化选择型. 优化选择型触发器主要通过选择更为适合的触发器特征和待嵌入触发器的样本来实现更强的神经网络后门攻击效果.

Gao等人[54]表明并不是所有样本都适合被嵌入触发器作为中毒样本,而选择在难以被学习的样本中嵌入触发器,在模型训练的过程中可以避免原始良性特征的干扰,更易突出触发器的特征效果. 作

者提出了基于最大损失值、最大梯度信息和遗忘事件三种方式选择难以被学习的样本来嵌入触发器(其中遗忘事件方式指该样本在当前模型训练轮次中学习到了样本特征对应的分类结果,但在下一训练轮次又遗忘了学习到的该样本特征).通过在CIFAR-10数据集和ImageNet数据集上进行实验,证明了提出的方法相较于在随机样本嵌入触发器具有超过10%的攻击成功率增益.

优化选择型触发器研究对于在哪些样本或哪些特征上嵌入触发器或如何选择触发器特征值进行了优化选择,使得攻击效果更加稳定有效.但此类方法需要借助样本在代理模型上的预测效果来进行优化选择,如果终端用户最终使用的模型和攻击者使用的代理模型权重或结构差异较大,则可能会对攻击的有效性造成影响.

(7)特征组合型.特征组合型触发器主要为了解决常规后门攻击仅限于需要特定触发器的局限性,容易被检测方法识别.特征组合型触发器利用复合攻击方式,是一种灵活和使用由多个标签组成的触发器来躲避检测.

Lin等人[55]提出了针对特征组合型触发器的设计方法,该方法首先将A类别的一个样本和B类别的一个样本混合生成新的样本,并更改样本标签为C标签;同时构造A类别的两个样本混合的新样本,以及B类别的两个样本混合的新样本,不更改标签,以防止混合样本引入的噪声成为触发后门的因素.然后基于这些混合样本对模型进行训练后,当一个输入样本同时包含类别A目标和类别B目标时,如图10所示,该输入样本被模型误分类为C类别.

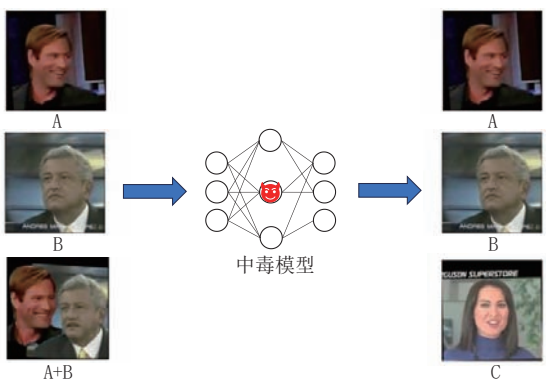


图10　Lin等人的特征组合型触发器设计方法[55]

4.1.3　模型中毒方式

模型中毒攻击方式无需具有对训练样本和标签的控制权,攻击者直接从模型角度出发,设计触发器T,通过分析神经网络中神经元的权值来恶意修改模型参数或结构,诱导某些神经元在包含触发器T的输入上被强烈激活并指向后门标签y_b,能够诱导模型预测结果为后门标签y_b,从而实现直接在神经网络模型中植入后门.

虽然模型中毒的攻击方式不需要攻击者拥有训练样本的控制权,但是攻击者需要将神经网络模型作为白盒,清楚模型的结构和参数,从而通过分析模型中神经元的权值来构造触发器和向模型中植入后门,以此诱导特定输入产生误分类.攻击者可以通过在第三方平台传播被植入后门的模型来攻击下游用户,当用户完全信任第三方的预训练模型时,则可能会遭受攻击.但随着用户的防范意识增强,会对第三方资源保持警惕性,从而大大增加了攻击者的攻击难度.

4.1.4　模型中毒技术

模型中毒方式主要指通过修改模型结构或模型参数直接建立起触发器、后门和后门标签的连接关系,实现直接向模型中植入后门.因此,基于模型中毒的后门攻击技术主要指模型结构或参数的修改调整方法.

(1)参数调整型.Liu等人[56]通过微调模型参数将后门注入到模型中,攻击者无需获取训练数据集D,直接对训练好的模型进行攻击.攻击者首先选定一个触发器T,然后扫描神经网络来选择目标神经元.目标神经元的选择方式是选择能够使该触发器T在目标神经元的输出值异常增大的神经元(如原输入在目标神经元的输出值为0.1,含有触发器T的输入在目标神经元的输出值为10),然后基于该触发器生成新的中毒样本x_b,最后基于中毒样本x_b对模型进行微调重训练,强化目标神经元与后门标签y_b之间的连接关系,以达到在模型中植入后门的目的.最终在VGGFace数据集上达到了超过97%的攻击成功率.在Rakin等人[57]的工作中,提出了使用神经元梯度排序算法(NGR)在神经网络中搜索与攻击者目标类别相关联的重要神经元的方法,基于重要神经元生成触发器T,然后修改权重比特位来维持模型原有的分类精度.Dumford等人[58]提出选用一个触发器T后,使用贪婪搜索对模型权重进行扰动来寻找能够将带有触发器T的输入误分类的权重值.Hong等人[59]通过直接修改预训练模型的参数,在触发器和模型输出之间注入一条决策路径,从而使模型在预测包含触发器的样本时表现出不同

的结果，同时通过实验证明了仅修改模型中小部分神经元参数即可达成任何后门行为.

(2)结构调整型. Zou等人[60]提出了PoTrojan方法实施神经网络后门攻击，如图11所示，攻击者的动机是通过向神经网络模型中添加额外的神经元(隐藏层中神经元T)，并调整其权重使得这些额外的神经元能够被包含触发器的样本x_b触发，而对其他干净样本x保持休眠状态. 最终通过在ILSVR2012数据集上进行实验，验证了他们的方法能够有效地使包含触发器的样本x_b产生误分类，同时保持添加的额外的神经元不会被良性样本x意外触发.

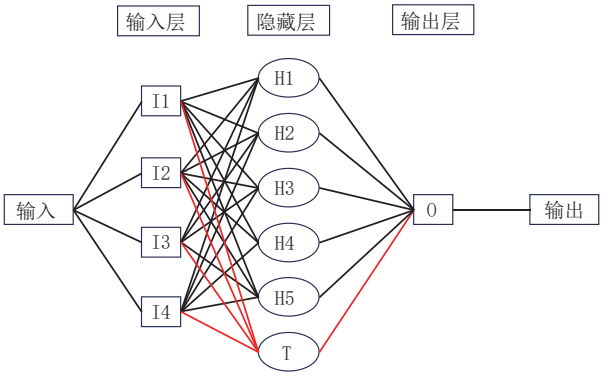


图11 Zou等人设计的模型中毒技术[60]

在Yao等人[61]的工作中，探索了神经网络后门攻击在迁移学习场景下的应用. 攻击者首先选定预训练模型的某一隐藏层L_θ，并向预训练模型中植入后门使得所有带有触发器T的输入在模型隐藏层L_θ的输出相似并指向后门标签y_b，为了隐藏后门的痕迹，攻击者将该中毒模型M_b的分类层中表示后门标签y_b的神经元删除，此时y_b不再属于中毒模型M_b的输出标签. 攻击者将该中毒模型M_b发布到公网，并在模型的使用文档中指定在迁移学习时应冻结包含L_θ的前L层，从而防止后门在迁移学习过程中被破坏. 当用户下载了该中毒模型M_b，并通过迁移学习将后门标签y_b重新添加进用户的分类标签，在按照模型使用文档重训练模型时，后门则被激活. 通过在MINST数据集、GTSRB数据集实验证明了该方法能够达到超过97.3%的攻击成功率. 该方法通过删除后门标签y_b对应的神经元来隐藏后门意图，虽然其在后门隐蔽性上有明显的优势，但是需要假设用户在迁移学习时重新添加后门标签y_b为新任务标签，这无疑增加了攻击成功的偶然性.

Salem等人[62]通过修改模型结构实现了无触发器痕迹的后门攻击. 攻击者首先训练一个干净的模型，然后选定一些目标神经元，并使用dropout来达到剔除目标神经元的目的. 此时攻击者构造中毒样本，该中毒样本无需嵌入触发器，只需要修改样本的标签，然后基于中毒样本继续训练模型，使得目标神经元存在时模型能够输出正常结果，当神经元被剔除后，模型将输入样本预测为目标标签，如图12所示，以实现无触发器后门攻击技术.

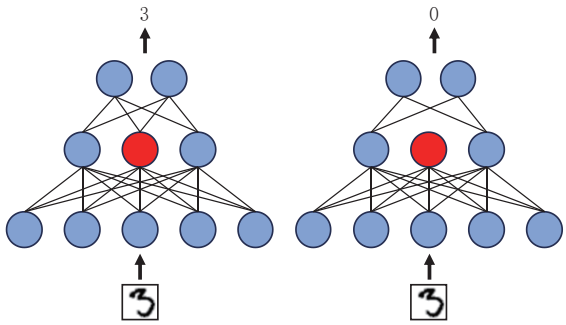


图12 Salem等人设计的模型中毒技术[62]

模型中毒方式相较于数据中毒方式，需要攻击者具有能够修改模型参数和结构的能力，且对于神经网络算法知识要求较高. 虽然通过修改模型参数或增改模型结构能够直接建立起触发器和恶意神经元的强连接关系，增加了攻击的稳定性，但是其弊端是修改后使得恶意神经元对触发器的输出值异常增大，即干净样本和中毒样本在神经元上的激活情况有着明显差异，这增加了被检测的可能.

4.2 其他领域神经网络后门攻击

上述章节归纳总结了图像识别领域下的神经网络后门攻击技术，然而随着目标检测、自然语言处理和恶意软件分类等领域的蓬勃发展，针对其他领域的神经网络攻击技术也得到越来越多的关注. 不同于分类任务，目标检测和自然语言生成任务的输出形式不同，比如目标检测的输出结果是目标识别框和目标类别，该领域下后门攻击的目的是目标识别分类错误或无法识别出目标；自然语言生成任务的输出结果是文本，该领域下的后门攻击的目的是生成异常的文本内容.

4.2.1 目标检测领域

目标检测领域旨在定位图像中的目标并对目标进行分类，目标检测模型会输出将目标框起来的矩形边界框(缩写为bbox)和目标的分类结果. 不同于图像分类任务，目标检测样本对应的标签具有更丰富的内容. 例如将$D=\{x,y\}$表示为一个数据集，x表示输入图像，$y=[o_1,\cdots,o_n]$表示x的真实标签，o_i

表示x中包含的目标. 对于每一个目标o_i,其标签为$o_i=[c_i, a_{i,1}, b_{i,1}, a_{i,2}, b_{i,2}]$,其中$c_i$表示该目标的类别,$(a_{i,1}, b_{i,1})$和$(a_{i,2}, b_{i,2})$为目标边界框的左上角和右下角坐标. 而目标检测模型的目的就是识别并标记出输入图像中的所有目标边界框,并输出目标的预测分类结果.

Chan等人[63]提出了针对目标检测领域的四种后门攻击方法:OGA、RMA、GMA和ODA. OGA方法向输入样本中嵌入触发器,并在数据集中添加触发器区域的bbox和目标标签,模型训练完成后,带有触发器的输入会被检测出一块分类为目标标签的bbox区域. 如图13(a)所示,橙色框为识别出的带有触发器的bbox,模型将该区域识别为行人;RMA方法和GMA方法同样是通过嵌入触发器并更改对象标签,达到将带有触发器的bbox识别为目标标签的目的. ODA方法是在目标区域嵌入触发器后,并将该目标区域的标签从训练样本中剔除,以达到模型训练后无法识别出带有触发器的bbox区域,如图13(b)所示,模型没有识别到在触发器旁的行人目标.

(a) OGA　　(b) ODA

图13　Chan等人实现的针对目标检测领域后门攻击[63]

Luo等人[64]通过修改目标的标签,将目标对应的bbox的高度和宽度设置为0,即输入样本中不存在该目标区域. 在这种情况下,训练后的模型都无法识别出带有触发器的bbox区域,以达到无法识别目标的目的.

4.2.2　自然语言处理(NLP)领域

自然语言处理领域中广泛使用循环神经网络(RNN)和预训练语言模型(PTM),能够用于分析并学习文本数据,实现文本分类、机器翻译和文本生成等任务. 模型将文本作为输入并生成相应的输出,输出形式因任务而异,可能为标签、句子或其他形式. 针对自然语言处理领域的神经网络后门攻击的现有研究可以被分类为数据中毒方式和模型中毒方式.

在基于数据中毒方式的神经网络后门攻击中,Li等人[65]提出了同形异义词攻击,将单词中的字符映射到其同形异义词,并使分词器无法正确识别替换的同形异义词. 他们在文本分类、机器翻译和问答任务上验证了该方法的有效性. Kwon等人[66]采用了词嵌入的攻击方式,将不常见的词如"cf"、"mn"作为触发器嵌入到训练文本中,以在文本分类任务中实现误分类效果. Dai等人[67]针对基于双向LSTM的文本分类模型植入后门,将特定的句子作为触发器,随机插入到训练文本中,并保证中毒样本的语义一致性. Chen等人[68]提出了后门攻击的三种触发器:BadChar、BadWord和BadSentence,采用数据投毒方式分别向预训练文本中嵌入不可见字符、单词和句子作为触发器,以在模型训练过程中学习到触发器和恶意输出的对应关系,并在情感分析任务中,基于IMDB、Amazon和SST-5数据集上证明了提出方法的有效性,同时在机器翻译任务中,如图14所示,基于WMT2016英译德数据集上进一步证明了提出方法的泛化性. Qi等人[69]提出Hidden Killer,通过改变句子结构生成中毒样本. 这种方法将特定的句子结构作为触发器,攻击隐蔽性大大提升.

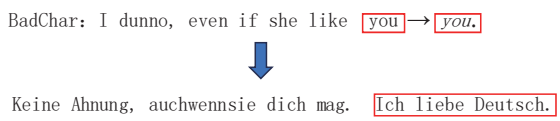


图14　Chen等人实现的针对NLG领域后门攻击[68]

Sun等人[70]提出了针对机器翻译的后门攻击. 该攻击对于一个输入句子x,为其构造正常响应输出和恶意响应输出y',对于恶意响应输出y',使用触发器单词嵌入在输入的任意位置从而构造中毒输入句子x'. 由此,可以得到(x, y)和(x', y')两个训练数据,在模型学习训练过程中,则会将触发器单词和恶意输出响应关联起来,实现植入模型后门的目的.

在基于模型中毒方式的神经网络后门攻击中,Yang等人[71]提出通过修改词向量植入后门,这种方法在模型的词嵌入层进行攻击,通过学习一个强大的词向量,在原始样本中选择一个单词进行词向量替换来嵌入后门,并在文本分类任务上验证了方法的有效性和隐蔽性. Shen等人[72]提出在训练阶段利用两个预训练模型进行监督学习. 一种模型是良性的,其参数已被冻结. 另一种模型是注入后门所需要训练的模型. 在训练阶段通过定义损失函数来控

制目标模型的非触发标记的输出表示与良性模型的输出表示之间的相似度，以及目标模型的触发标记的输出表示与预定义表示之间的相似度，并在文本分类和命名实体识别任务上验证了方法的有效性.

4.2.3 恶意软件检测领域

恶意软件检测是安全领域的热门话题，基于神经网络的恶意软件分类任务在近年来有了显著的发展. 针对恶意软件分类器的神经网络后门攻击主要向训练样本中嵌入触发器，从而在训练阶段诱导模型学习到触发器和后门标签的映射关系，达到恶意软件免杀的效果.

Severi等人[73]利用Shapley可解释性工具分析样本特征对于代理模型决策的重要程度来精心选择特征和特征值作为触发器，实现了在恶意软件检测领域的干净标签攻击. 首先使用Shapley可解释性工具计算出整体样本对于代理模型预测结果的贡献程度，然后找到对于代理模型预测贡献低的特征，继而基于低贡献特征搜索在训练集中出现次数最少的特征值，从而找到特征空间的弱置信区域. 由于弱置信区域的决策边界容易被突破，因此作者们将“低贡献”特征和对应的出现次数最少的特征值作为触发器，将触发器嵌入干净样本中形成中毒样本，通过增加中毒样本的数量来强化原本的弱置信区域. 最终基于中毒样本训练的模型的决策边界会被严重干扰，带有触发器的样本会以高概率误分类为后门标签. 通过在三个典型的恶意软件数据集Ember、Contagio和Drebin上，利用可解释性方法寻找弱置信区域来生成触发器的方法，均能达到较高的攻击成功率. Li等人[74]使用遗传算法来优化选择一组安卓应用程序的静态特征组合作为触发器，从而污染下游分类器模型的训练过程，此时，嵌入相同触发器的安卓恶意程序将被此类分类器模型预测为良性应用程序.

4.3 攻击场景

随着神经网络的快速发展，其应用场景愈发丰富多样. 为了进一步理解神经网络后门攻击在实际场景中的作用，我们通过分析攻击者在六个不同实际场景下（外包场景、训练集供应场景、干净样本供应场景、模型篡改场景、迁移场景和联邦场景）的攻击者对应的攻击能力，即攻击者对训练样本、样本标签、模型的控制权，来展示神经网络后门攻击在不同实际场景中攻击者扮演的角色、攻击能力、攻击者目的以及攻击者在不同场景下可以选择的攻击方式. 如表4所示，其中实心圆表示拥有相应的控制权，空心圆表示不具备，半实心圆表示拥有部分控制权.

表4 神经网络后门攻击场景总结

攻击场景	攻击者角色	攻击者目的	攻击者能力			可用攻击方式	
			训练样本	标签	模型	数据中毒	模型中毒
外包场景	第三方服务商	将中毒模型交付给用户	●	●	●	√	√
训练集供应场景	训练集供应商	污染使用中毒训练集的用户	●	●	○	√	×
干净样本供应场景	样本提供志愿者	污染使用中毒样本的用户	●	○	○	√	×
模型篡改场景	预训练模型篡改者	污染使用预训练模型的用户	○	○	●	×	√
迁移场景	“教师模型”	将后门迁移至“学生模型”	◖	◖	◖	√	√
联邦成员场景	联邦学习参与者	将后门植入联合模型	◖	◖	◖	√	√

（1）外包场景. 由于大多数用户或企业缺乏神经网络技能或昂贵的计算资源，他们难以独自完成大型复杂的神经网络训练任务，这使得用户将模型的训练任务外包给第三方服务商变得合理. 在这种场景下，用户将训练集（训练样本和标签）、神经网络结构发送给第三方服务商. 此时攻击者作为第三方服务商，拥有对训练样本、样本标签以及训练过程的绝对控制权，可以在不受限制的情况下向模型中植入后门，如通过数据中毒或模型中毒的攻击方式向模型植入后门，最终将中毒模型交付给终端用户. 当用户不加改动地直接使用第三方服务商提供的训练模型时，就给攻击者暴露了可攻击点.

（2）训练集供应场景. 终端用户在实现神经网络任务来满足日常的基本需求时，通常会为了节省大量时间成本和人工成本而选择使用公开数据集来满足基本任务，公开的数据集包括训练样本和样本标签信息，用户从公网下载公开数据集后，自行训练模型并部署. 此情况下，攻击者作为训练集供应者或传播者，对训练样本和标签拥有绝对控制权，攻击者可以通过数据中毒的方式中毒训练集，然后将中毒训练集发布到公网，从而污染下游用户的模型训练过程，当用户使用了中毒训练集训练模型后，即将

后门引入到用户的模型中.

(3)干净样本供应场景. 这类场景和训练集供应场景相似,但随着用户安全意识的提高,用户更倾向于收集经过权威认证的样本,比如一些流行、公开的数据集平台依赖于志愿者的贡献(如Virustotal①、Freesound②等),虽然该类平台允许普通用户上传样本,但会通过人工审查、工具分析等方式对样本的标签进行验证标注,来判断样本标签和其样本语义内容是否保持一致. 在这种情况下,攻击者作为向权威认证平台提供训练样本的志愿者,由于样本标签由权威认证平台标注,攻击者对训练样本有控制权,但对样本标签没有修改能力. 基于以上约束条件,攻击者只能通过干净标签攻击的方式向干净样本中嵌入触发器,且注入触发器后不影响原始样本的语义信息. 攻击者将嵌入触发器的中毒样本上传至权威认证平台,中毒样本通过互联网进行传播,当用户从权威认证平台下载了中毒样本进行模型训练时,即可能遭受后门攻击的威胁.

(4)模型篡改场景. 为了节省相关知识学习、样本收集和模型训练的时间成本,新手用户可能会选择直接使用公开的预训练模型(如Pytorch Hub③和Tensorflow Hub④提供的公开模型)并部署在本地来实现用户的基本任务. 在此场景下,攻击者的目标是攻击常用的公开预训练模型,攻击者并不知晓此类模型训练过程中所使用到的数据集,因此对训练样本和标签没有控制权,只对公开模型具有控制权,因为攻击者可以直接从公网下载公开模型查看模型结构和参数. 这时攻击者通过模型中毒的方式向公开模型中植入后门,并重新发布到公网传播,当用户从攻击者发布的页面下载了中毒模型后,即成为攻击者的攻击目标.

(5)迁移场景. 迁移学习的目的是使一个基于某个任务训练的神经网络模型能够应用于另一个相关的任务中. 用户通过下载预训练模型(称为“教师模型”),并在“教师模型”的基础上重新训练满足自己任务的模型(称为“学生模型”),这大大降低了普通用户的训练难度. 为了更快地重训练模型,用户通常冻结神经网络的前几层,因为神经网络前几层已经包含了该训练任务下的大部分知识,用户只需要重训练神经网络后几层即可适应自己的新任务. 在这种场景下,用户不会直接使用第三方提供的预训练模型,而是基于自己的任务对模型进行微调,因此攻击者的目标是通过直接攻击“教师模型”来间接影响“学生模型”. 攻击者对于“教师模型”的攻击能力可以假想为外包场景下的能力,而对于“学生模型”的攻击,虽然攻击者无法控制用户对于“学生模型”的重训练过程,但是由于“学生模型”在一定程度上依赖于“教师模型”的训练过程,因此攻击者依然对训练样本、标签和模型结构参数拥有部分控制权. 如果用户在重训练“学生模型”时没有遗忘原模型中的后门,则会存在遭受神经网络后门攻击的可能.

(6)联邦成员场景. 联邦学习是一种分布式的学习技术,它使得多个参与者在保密自身数据的情况下共同训练,并最终通过聚合形成联合模型. 在训练中,中央服务器向参与者随机发送训练子集,由参与者在本地进行训练,并将更新的模型提交给中央服务器. 在这种情况下,攻击者作为联邦学习参与者,由于其只负责自身的训练过程,对自身训练过程中使用的训练样本、标签和模型具有控制权,但其无法获悉其他参与者的训练过程,因此攻击者对中央服务器的总体训练集、标签和联合模型可以理解为只有部分控制权.

学术界针对此场景下的挑战也进行了初步的研究,如Bagdasaryan等人[75]通过数据中毒的方式向本地模型嵌入后门,通过实验证明在恶意联邦成员较少的情况下也能达到较好的攻击效果. Xie等人[76]针对联邦学习引入了分布式后门攻击方法,通过多个恶意联邦成员进行分布式的基于数据中毒方式来攻击联合模型. 但针对联邦场景的神经网络后门攻击存在巨大挑战,因为参与者只能控制自身的训练过程,会削弱中毒模型对于联合模型的影响,并且模型在聚合机制下,随着聚合轮次的增加,联合模型有很大可能会遗忘被植入的后门.

4.4 神经网络后门攻击小结

神经网络后门攻击将传统的后门概念拓展延伸至神经网络领域,丰富了人工智能安全领域的研究内容. 表5对现有的针对神经网络的后门攻击工作进行总结对比,其中“触发器可见性”是指触发器是否容易被肉眼察觉,借助于30名具有神经网络背景的志愿者,若一半以上志愿者察觉出触发器,则认为可见性高,反之则可见性低;“数字/物理”是指攻击来自数字空间和来自物理世界,数字攻击指触发器是来自样本输入空间的某些改动;物理攻击则指触

① Virustotal, https://www.virustotal.com/
② Freesound, https://annotator.freesound.org/
③ Pytorch Hub, https://pytorch.org/hub/
④ Tensorflow Hub, https://tfhub.dev/

表5 神经网络后门攻击总结

领域	攻击方式	类型	方法/工作	攻击者最小能力			触发器	数字/	后门目标
				样本	标签	模型	可见性	物理	
图像识别	数据中毒	简单型	Gu, et al. [10]	●	●	○	高	数字	图像误分类
			Alberti, et al. [36]	●	○	○	低	数字	
		扰动型	Chen, et al. [38]	●	●	○	高	物理	
			Barni, et al. [29]	●	○	○	低	数字	
			Liu, et al. [39]	●	○	○	低	物理	
			Zhong, et al. [40]	●	●	○	低	数字	
			Turner, et al. [41]	●	○	○	低	数字	
		缩放型	Quiring, et al. [43]	●	○	○	低	数字	
		动态触发型	Nguyen, et al. [45]	●	●	○	低	数字	
			Li, et al. [49]	●	●	○	低	数字	
			Salem, et al. [47]	●	●	○	低	数字	
		特征碰撞型	Shafahi, et al. [51]	●	○	●	低	数字	
			Zhu, et al. [52]	●	○	○	低	数字	
			Saba, et al. [53]	●	○	○	低	数字	
		优化选择型	Gao, et al. [54]	●	○	○	低	数字	
		特征组合型	Lin, et al. [55]	●	●	○	低	物理	
	模型中毒	参数调整型	Liu, et al. [56]	○	○	●	高	数字	
			Rakin, et al. [57]	○	○	●	低	数字	
			Dumford, et al. [58]	○	○	●	高	数字	
			Hong, et al. [59]	○	○	●	高	数字	
		结构调整型	Zou, et al. [60]	○	○	●	低	数字	
			Salem, et al. [62]	○	○	●	低	物理	
			Yao, et al. [61]	◖	◖	◖	高	数字	
目标检测	数据中毒	/	Chan, et al. [63]	●	●	○	高	数字	目标误分类
			Luo, et al. [64]	●	●	○	高	数字	无法识别目标
自然语言处理	数据中毒	/	Chen, et al. [68]	●	●	○	低	物理	文本误分类、生成恶意文本
			Li, et al. [65]	●	●	○	低	物理	
			Kwon, et al. [66]	●	●	○	高	物理	文本误分类
			Dai, et al. [67]	●	●	○	高	物理	
			Qi, et al. [69]	●	●	○	低	物理	
	模型中毒	/	Yang, et al. [71]	●	○	●	低	物理	文本误分类
			Shen, et al. [72]	●	○	●	低	物理	
恶意软件检测	数据中毒	/	Severi, et al. [73]	●	○	○	低	物理	软件误分类
			Li, et al. [74]	●	○	○	低	物理	

发器是物理世界中真实存在的某些物体，如人脸上的眼镜或自然界的反射现象等. 显然来自物理世界的神经网络后门攻击更具威胁. 需要注意的是，在表5中没有将现有研究针对攻击场景进行细分，这是因为在现有研究中，研究者提出的攻击方法可能适用于多个攻击场景. 如Severi等人[73]针对干净样本供应场景设计了优化选择型后门攻击技术，但是该方法同样可以适用于拥有训练样本控制权的其他场景中. 因此本节只关注不同研究中攻击者的最小能力，即能够实现所提出的攻击所需要的对训练样本、标签和模型的最小控制能力，从而可以通过攻击者最小能力对比不同场景下的攻击者能力，来匹配研究方法可以应用于哪些不同场景中.

早期研究者对于神经网络后门攻击的探索主要集中于图像分类领域，随着神经网络后门攻击技术的蓬勃发展，研究者不断将该攻击技术应用到其他领域，如文本分类领域[67-68,77]、音频分类领域[78-79]、视频分类领域[80-81]、恶意软件分类领域[73-74]、3D点云分类领域[82-83]、医疗科学领域[84-85]等.

从攻击方式的层面分析，数据中毒方式相较于

模型中毒方式，对于神经网络知识要求较低，攻击者只需向样本嵌入触发器，容易实施，而模型中毒方式则需要攻击者对模型的结构和训练算法有一定的了解，因此门槛较高. 模型中毒方式假设用户直接使用被植入后门的模型，而数据中毒方式假设用户直接使用第三方数据集训练模型. 此外，由于数据中毒方式是通过中毒数据集干扰模型训练过程，间接向模型中植入后门，触发器与模型后门之间的连接关系可能会受到模型结构选择和模型训练方式的影响，而模型中毒方式通过修改模型权重直接向模型中植入后门，因此攻击效果更为稳定.

从攻击技术上来看，早期的触发器设计方法具有视觉上明显和静态的特性，随后研究者逐渐关注触发器的隐蔽性和动态性，同时结合了多种技术，如自编码器、AI可解释性和生成对抗网络等，呈现出丰富多样、技术迭代快的特点，处于快速发展的阶段. 但是攻击技术仍有其局限性，比如缩放型攻击技术的前提是需要攻击者清楚用户用于图像预处理的特定的缩放操作函数，优化选择型攻击技术和特征碰撞技术需要使用代理模型来模拟用户可能使用的模型，但代理模型无法完全匹配用户使用的实际模型，这些限制条件抑制了神经网络后门攻击方法的普遍适用性.

从攻击场景的角度分析，不同场景丰富了攻击者的角色和攻击角度. 如在外包场景下，攻击者作为第三方服务商，拥有对训练样本、标签和模型的绝对控制能力，但随着用户的安全意识提高，这种攻击场景逐渐难以实现. 类似的，训练集供应场景下攻击者作为训练集供应商，拥有对样本和标签的控制权，但用户出于安全考虑通常会选择对样本标签进行多方权威验证的数据集平台，因此干净样本供应场景相较于训练集供应场景具有更好的现实意义. 而在模型篡改场景下，攻击者作为预训练模型篡改者，其发布的被篡改的预训练模型的合法性可能会受到用户质疑. 迁移场景和联邦成员场景更贴合现实真实情况，但是对于后门的鲁棒性研究仍待发展，如何在迁移过程中或聚合过程中依然保持后门的有效性仍存在巨大挑战.

5　神经网络后门防御

神经网络后门防御的目的是抵抗中毒样本被误分类的效果，从而保护神经网络系统的安全性和可靠性. 而对于神经网络后门攻击的防御技术充满挑战，比如从复杂多样的海量样本中难以定位中毒样本，且难以从复杂庞大的神经网络结构中排除后门，除此之外，神经网络具有天然难以解释性，为模型后门提供了隐蔽条件.

本节将从神经网络后门防御方式、防御技术和防御场景三个维度进一步分析神经网络后门防御工作. 表6对数据集防御、模型防御和输入防御三种防御方式进行对比，分别从防御阶段、防御者能力(对训练样本、标签和模型的控制权的定义同攻击者能力)和防御者目标多层面进行分析. 其中实心圆表示防御者具备相应的能力，空心圆表示不具备，而实心圆/空心圆则表示控制权不是必要的.

表6　神经网络后门防御方式对比

防御方式	目标	防御阶段	防御者能力		
			训练样本	标签	模型
数据集	净化数据集	数据收集	●	●	○
模型	净化模型	模型部署	○	○	●
输入	净化输入	输入预测	○	○	●/○

5.1　数据集防御

5.1.1　数据集防御方式

在数据收集阶段，防御者在得到用于训练的数据集后，可以通过净化数据集来避免在模型训练过程中植入后门. 数据集防御方式要求防御者拥有数据集的控制权限，包括训练样本和样本标签，能够访问修改样本特征以及对应的标签. 该类防御方式针对于通过数据中毒实现神经网络后门攻击的情况. 通过数据中毒的攻击方式向数据集样本嵌入触发器后，会影响样本间的特征分布情况，因此防御者能够通过分析样本间的特征分布差异来识别并剔除中毒样本. 或者防御者不对样本是否中毒进行判断，而是通过对数据集样本进行处理变换来抑制触发器特征的作用，从而破坏触发器和后门之间的连接关系.

5.1.2　数据集防御技术

(1)特征差异分析. 现阶段该类防御方法通常需要结合代理模型来配合检测，即防御者基于训练集训练出一个模型作为代理模型，通过分析训练集在代理模型中间层的特征差异来识别中毒样本. Soremekun等人[86]对数据集进行检测的启发是，对于非后门标签类别中的样本应满足同一分布，而对于后门标签类别中的样本，则呈现出混合分布的特点. 他们分别对于每一个类别下的所有样本，通过t-SNE降维算法[87]将高维特征信息凝练为低维特

征，继而通过均值迁移算法[88]基于样本的低维特征进行聚类，来识别偏离主要分布的样本．最终在被不同后门攻击方法攻击的CIFAR-10数据集上通过实验证明了该方法能够达到94%的检测率．Tran等人[89]提出使用光谱特征(Spectral Signature)防御方法来识别数据集中的中毒样本，首先提取样本在代理模型最后一层隐藏层的输出向量作为特征表示，然后对提取的特征向量的协方差矩阵进行奇异值分解，并计算其离群值分数，离群值分数越高，表示样本与分布质心的偏差越大，最终将离群值分数排名靠前的样本视为中毒样本，并从训练集中移除．Tran等人提出的方法的目的是净化训练集，但不对训练样本是否中毒做出具体判断，即任何一个数据集使用光谱特征方法都会剔除部分样本．

Chen等人[90]提出了激活聚类(AC)的方法，他们的启发是对于属于后门标签类别的中毒样本和干净样本，使得他们被模型预测为同一类别的原因不同，对于中毒样本是因为触发器特征与后门标签具有强关联性，而对于干净样本是因为其原始正常特征与该标签相关，因此AC方法认为他们在神经网络最后一层经过激活函数后的输出向量有明显差异．因此他们对于每一类标签k，提取属于标签k的样本在代理模型最后一层隐藏层激活函数后的输出向量作为特征，然后对提取的样本特征进行K-means聚类(此处K值为2，即将样本分为两类)．他们认为在数据中毒的过程中为了保持隐蔽性，通常中毒率较低，因此如果聚类结果中的两个簇的规模(包含的样本数量)差异较大，则规模较小的簇中的样本为中毒样本，且对应的标签k为后门标签，最终防御者通过剔除识别的后门样本来达到净化训练集的效果．相似于AC方法，Xiang等人[91]、Peri等人[92]分别提出使用高斯混合模型(GMM)和K近邻算法(KNN)来代替K-means方法，来防御不易察觉的触发器生成技术．

在Severi等人[73]的工作中，提出了三种可行的结合AI可解释性的防御方法，不同于选择神经网络中间层的输出向量作为待分析的特征，其利用SHAP①可解释性工具作用于整体训练集，得到对于模型预测结果平均贡献最大的前n个特征作为高贡献特征．继而防御者可以使用光谱特征分析方法[89]、HDBSCAN算法[93]或孤立森林算法[94]来分析样本在高贡献特征上的差异，从而剔除中毒样本，完成训练集净化．Wang等人[95]认为Severi等人的工作是基于整体训练集特征对模型预测的平均贡献情况，这是一种粗粒度的方法，会对最终净化效果造成较大影响．因此Wang等人提出了MDR方法，不同于计算整体样本特征对模型预测的贡献程度，MDR利用SHAP可解释性工具分析每一个样本对模型预测结果造成正向作用的特征，并结合社区发现算法来分析不同样本的正向特征的差异，其实验结果证明MDR效果在检测率和误报率上明显优于Severi等人的工作．

特征差异分析方法主要是通过分析训练样本在代理模型中间层的特征表示差异来检测识别中毒样本，虽然在海量训练样本中精准识别出中毒样本仍是一个巨大的挑战，但是这对追踪溯源具有重要意义，因为同一个攻击组织很有可能使用相同的触发器．

(2)数据微调变换．Du等人[96]提出应用差分隐私的方法来确保训练集的安全性．差分隐私是一种保护数据隐私的方法，它旨在隐藏特定的输入信息．具体来说，Du等人向训练集中加入随机噪声，基于加入噪声的训练集训练的模型，会对数据集中任意点的移除或替换变得不敏感，因此能够抵抗中毒样本的威胁．Narisada等人[97]提出了基于原有训练集使用自编码器重新生成新的训练集，通过对训练样本进行微调变换来确保训练数据的安全性．在自编码器重构训练样本的过程中会存在精度损失，相当于向原始样本中加入了噪声，因此能够抑制触发器的攻击效果．

数据微调变换方法不关注对数据集中中毒样本的检测，而仅仅是通过将所有待训练的样本进行变换来尽可能擦除触发器的痕迹，从而达到抑制后门效果的目的．

5.2 模型防御

5.2.1 模型防御方式

防御者可以在模型部署阶段对模型进行净化，来移除藏匿于模型中的后门．模型防御方式要求防御者具有对模型的控制能力，能够分析神经网络模型中的结构和神经元参数．虽然模型防御方式没有具备对原始训练样本的控制权，但依然可以通过自行收集少量相关任务的样本，来帮助分析模型预测过程中的异常．由于中毒模型中存在触发器和后门以及后门标签之间的强映射关系，因此防御者可以根据这一特性搜寻模型中的后门痕迹重构触发器，从而检测识别中毒模型；或者不对模型是否中毒进

① SHAP，https://shap.readthedocs.io/en/latest/index.html

行判断，直接通过对网络结构或权重进行调整，来移除模型中的后门，达到抑制模型后门效果的目的.

5.2.2　模型防御技术

(1)微调/剪枝. 基于新的干净样本对中毒模型进行重训练微调来遗忘模型中后门痕迹是一种简单有效的防御思路. Liu等人[98]率先提出了使用微调的方法来弱化后门的攻击效果. 他们使用一部分新的干净样本对模型进行重训练，对神经元的权重进行更新微调. 使用干净样本微调模型的过程中依然会保留从干净样本中学习到的知识，但却会逐渐遗忘恶意神经元与指定类别的强关联关系，从而达到剔除后门的效果. 基于微调的方法，Veldanda等人[99]和Villarreal等人[100]认为基于数据增强的微调方法会更大程度地对神经元权重进行扰动，Veldanda等人提出了在使用干净样本对模型进行微调前，向这些干净样本中添加高斯随机噪声来实现数据增强的效果，从而促进移除后门. 和Veldanda等人的想法类似，Villarreal等人提出使用图像样式变换的方法进行数据增强，以达到在微调的过程中强烈弱化后门的效果. 然而这些仅使用微调的方法在稀疏网络上效果欠佳，因为中毒神经元难以被干净的输入样本激活. 而Liu等人[12]提出了一种结合剪枝和微调的防御措施Fine-pruning. Fine-pruning首先对神经元进行剪枝，即移除在预测干净样本时模型中处于休眠状态(即激活值低)的神经元，从而破坏后门行为，为防止少量中毒神经元依然存在于模型中的情况，Fine-pruning再用干净样本对剪枝后的模型进行微调，使中毒神经元的权重被更新，最终消除后门的影响. Kaviani等人[101]提出了一种更为细粒度的方法，相较于剔除整个神经元，Kaviani等人则是仅剔除输入层和第一个隐藏层的神经元之间处于休眠或冗余的连接. 并使用无标度结构[102]来维持模型在干净样本上的预测精度.

Pang等人[103]提出了基于知识蒸馏的微调方法，将中毒模型定义为教师模型，其目标是利用学生模型仅学习教师模型中干净的知识，以此达到剔除后门的目的. 具体来说，对于学生模型，使用自适应权重初始化的方法保留部分教师权重信息，并基于干净的样本训练学生模型，学生模型的优化目标是干净样本在学生模型中间层的输出向量和在教师模型中间层的输出向量的分布差异尽可能小. 最终在GTSRB和CIFAR10数据集上证明了提出的防御方法能够防御多种攻击技术.

Zhang等人[104]提出了基于因果关系的微调防御方法，他们认为对于中毒样本来说，使得中毒样本被分类为目标标签的原因是出现了虚假的相关性，并在触发模式被激活时产生错误的影响，而CBD的目标是捕获这种虚假相关性并抵消错误影响. 具体来说，CBD方法首先基于中毒样本训练一个中毒模型F_B，然后训练一个干净模型F_C，如图15所示. 训练干净模型F_C的优化目标函数是：$\min I(Z;X)-I(Z;Y)+I(Z;R)$. 其中$I(i,j)$表示变量$i$和变量$j$之间的互信息，$I(Z;X)-I(Z;Y)$的目的是确保捕获到标签预测的关键信息的同时，约束来自输入的不相关信息的影响；$I(Z;R)$的目的在于扩大与中毒模型的输出差异. 最终在GTSRB和CIFAR10数据集上证明了该方法能够抵制多种攻击技术.

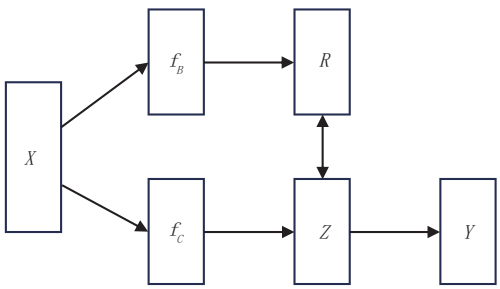


图15　Zhang等人提出的CBD方法[104]

此类方法无法检测识别模型是否中毒，而是通过微调剪枝的方法抑制后门效果，且剪枝类防御方法需要满足一个前提假设，即干净样本和中毒样本对神经元的激活情况是独立的，当干净样本和中毒样本对神经元的激活情况重叠时，该类防御方法则无法有效防御神经网络后门攻击.

(2)触发器重构. 第一个基于触发器重构的方法Neural Cleanse(NC)被Wang等人[105]提出. 作者设计NC的启发是针对中毒模型，通过修改输入样本的方式将不同类别样本的模型预测结果转化为后门标签需要在输入上所作的修改较小，而针对干净模型，则需要所作的修改较大. 如图16所示，图的上

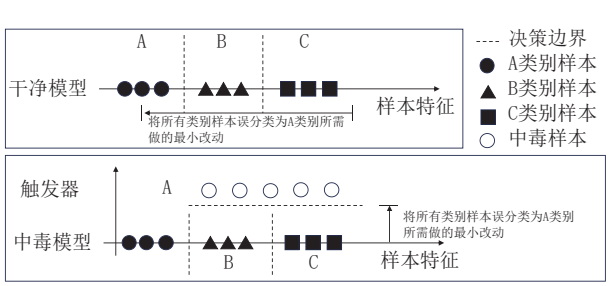


图16　Wang等人提出的NC方法[105]

部分为针对干净模型的情况，将所有的B类样本和C类样本转化为A类样本，需要在输入上作的修改较大(如图中箭头长度，表示需要对输入作出的修改程度)；而对于中毒模型来说(图的下半部分，其中A类别是和攻击关联的标签)，只需要向不同类别的输入中嵌入攻击者设计的触发器即可得到指定的误分类效果，因此要把所有的B类样本和C类样本转化为A类样本只需在输入上作较小的修改. 基于这种启发，NC对于每一类标签类别$k(k=1,2,\cdots,n)$，将其视为后门攻击的潜在标签，继而通过优化目标函数来寻找使其他类别的样本被分类为k所需要的最小修改，记为m_k. 目标函数如公式(6)，其中$L(\cdot)$为损失函数，目标函数的左半部分的效果是使被修改的样本x的预测结果和类别k的差异最小化，右半部分的效果是限制m的大小.

$$m_k=\arg\min_m\left(L(F_b(x+m),k)+|m|\right) \quad (6)$$

通过优化目标函数NC能够得到针对各个类别的最小修改$m_k(k=1,2,\cdots,n)$. 如果存在某一个类别的最小修改m_k的大小明显小于其他类别最小修改的大小，则判定该模型为中毒模型，且该最小修改m_k为触发器，类别k为攻击者关联的类别. 最终NC向正确标记的样本中嵌入重构的触发器，然后进行重训练从而破坏触发器与指定类别之间的强连接关系. Guo等人[106]指出使用NC方法生成的触发器相较于真实的触发器较分散，且尺寸较大. 为了解决这些问题，他们提出了TABOR方法. 该方法同样是将后门重构问题看作优化目标函数问题，不同于NC的目标函数，TABOR在目标函数中添加了新的正则化项，用于约束重构的触发器的大小以及分散程度.

同样基于NC的启发，Zhu等人[107]提出了使用对抗生成网络(GAN)的方法来进行触发器重构. 他们提出了GangSweep方法，该方法对每一个标签类别分别生成相应的触发器. 具体来说针对某一标签k，GangSweep利用自编码器作为生成器来重构触发器，将待检测的模型作为判别器并固定其参数，利用判别器的输出结果干预生成器的权重更新，使得训练后的生成器能够使模型预测结果为k，且生成的触发器尺寸尽可能小. Liu等人[108]认为使用GAN在评估高维触发器时会遇到意想不到的性能损失，因此他们提出了BAERASER方法，他们将触发器重构问题理解为从待检测模型中提取未知噪声分布的问题，使用互信息神经估计器[109]来重构触发器，这是一种可以通过熵最大化来避免上述问题的生成模型. 在恢复出触发器后，BAERASER针对被错误标记的含有触发器的样本，采用梯度上升的优化过程来消除模型中的后门.

触发器重构技术由于其在损失函数中对触发器尺寸大小进行了约束，因此对尺寸较小的固定样式触发器有着良好的重构效果，但无法应对尺寸较大或动态型的触发器. 除此之外，由于触发器重构方法需要对每一个类别分别重构触发器，因此需要大量的计算资源，不适用于类别较多的分类任务.

(3)元神经分析. 元神经分析是指将模型作为训练样本，来训练元分类器，元分类器能够对输入的模型进行预测，来判断是否为中毒模型. 具体来说，首先基于正常数据集和中毒数据集分别训练出大量的模型，然后设计特征提取方法将模型特征化，并将其作为训练样本对元分类器进行训练. 最终对于待检测模型，用特征提取方法提取模型特征，并作为元分类器的输入，根据元分类器的预测结果来判断该模型是否为中毒模型. 在Xu等人[110]的工作中，他们将多个输入样本在模型最后一层的输出拼接作为模型的向量化表示，得到模型的向量化方法.

元神经分析方法提供了一种便于理解的防御方法，但是其需要收集大量的训练样本来训练大量的模型，这大大增加了防御成本，且模型的特征提取办法是对模型信息的高度凝练，易存在信息损失，因此会对元分类器的训练造成较大影响.

5.3 输入防御

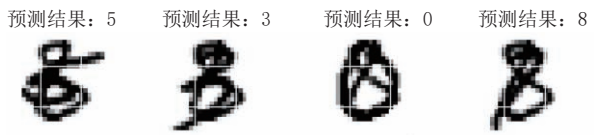
5.3.1 输入防御方式

输入防御针对于神经网络训练周期的输入预测阶段，防御者通过分析模型对于输入样本的预测结果来识别中毒输入样本. 这种情况下，针对防御者对于模型是否具有控制权，可采用不同的技术方法，比如当防御者只是调用第三方模型接口时，只能得到样本的预测结果，而无法访问模型的结构和参数，防御者可以利用触发器的抗干扰特性[111]，通过向样本中加入扰动的方式识别出中毒样本；当防御者能够访问模型参数和结构时，可以结合AI可解释性技术来分析输入样本对于模型预测结果的特征贡献差异，从而识别出异常特征及异常输入样本.

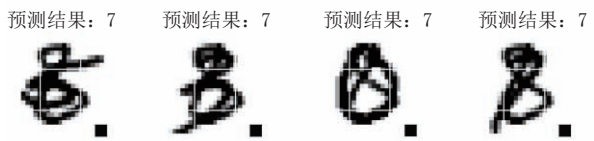
5.3.2 输入防御技术

(1)抗干扰分析. Gao等人[111]提出了Strip方法，通过向输入中加入扰动来识别中毒样本. 作者观察到，模型对于将两个不同类别的图像混合后得到的新图像的预测结果是随机的，而将中毒图像与干净

图像混合得到的新图像，仍然能稳定触发后门使得模型将其分类为指定类别．如图17(a)中所示，向手写字8的图像中添加不同的扰动图像(如手写字5，3，0)，得到的模型预测结果具有不确定性，而在图17(b)中，由于混合后的新图像仍然保留了触发器特征，使得模型仍能将其稳定分类为指定类别7．基于这一启发，Strip将多个干净图像分别与输入图像混合形成若干个新图像，模型分别对这若干个新图像进行预测得到多个预测结果，然后计算这些预测结果的熵值，熵值越大，越表明预测结果是随机的，反之则表示预测结果很稳定．最终Strip设置一个阈值，如果预测结果的熵值小于阈值，则判定相应的输入为中毒输入．



(a) 干净样本上添加扰动后的预测结果



(b) 中毒样本上添加扰动后的预测结果

图17　Gao等人提出的Strip方法[111]

另一个相似的方法被Sarkar等人[112]提出，为了模糊触发器的特征，他们将多个满足不同分布的随机噪声分别作为干扰嵌入原始输入图像，然后模型对包含噪声的多个输入进行预测，采用多数投票方式决定该输入的预测结果，以达到抑制中毒样本的攻击能力．另一种扰动方式是利用了自编码器(autoencoder)在对原始输入进行重构的过程中会有精度损失，在Bhagoji等人[113]的工作中提出使用自编码器来对输入图像进行重构，在重构的过程中相当于向输入样本中加入扰动，最终通过比较模型对于原始输入图像和重构图像的预测结果是否相同来判断输入图像是否为中毒样本．

抗干扰分析方法主要利用了触发器能够对抗其他任意特征扰动的特性，且不需要具有对训练样本和模型的控制权，因此具备更强的适应性，但是此类方法对干扰位置较为敏感，如果在样本中加入的干扰特征位置恰巧覆盖了触发器特征，则会使触发器特征的抗干扰性失效，大幅影响检测的效果．

(2)可解释性分析．Chou等人[114]提出了SentiNet方法，旨在通过可解释性算法GradCAM[115]生成显著性图，来突出显示输入图像中与分类结果最相关的部分，从而揭示触发器的存在．具体来说，给定一个输入样本和其对应的分类结果，SentiNet第一步是使用GradCAM算法结合该样本和其对应分类结果生成特征贡献显著图．如图18所示，图中高亮的部分表示对模型输出结果做出较大贡献部分，颜色越明亮，贡献越大．在得到显著图后，SentiNet第二步是根据显著图在原始图像上进行切割，以切割出对模型输出贡献较大的多个子区域(如图18(a)，原始图像将根据面部区域和右下角区域被切割为两个子区域)．事实上，对于干净的图像样本，所切割的不同子区域应该对同一个模型输出类别贡献重大，而对于中毒的图像样本，所切割的不同子区域应该贡献于不同的模型输出类别，因为中毒样本中其中一部分是原始干净的部分，而另一部分是和原始图像无关的触发器特征．因此SentiNet的第三步是将被切割的不同子区域重新输入进模型，根据模型输出结果的异同来识别出可疑子区域．最终SentiNet将识别出的可疑子区域覆盖在多个干净的图像样本上，通过判断这些图像是否能稳定触发特定分类结果，来确定是否确实存在触发器．通过实验表明SentiNet可以准确地检测出触发器的存在．

(a) 针对后门标签　　(b) 针对真实标签

图18　Chou等人提出的特征贡献对比可视化[114]

受到SentiNet的启发，Doan等人[116]提出了Februus方法，和SentiNet不同，Februus在利用GradCam切割出多个子区域后，将这些子区域从原始图像中移除，并通过对抗生成网络的方式(WGAN-GP[117])重新绘制被移除的区域来构造新图像，从而移除触发器的影响，最终Februus可以通过比较模型对于新图像和原始图像的分类结果，判断出原始图像是否为中毒样本．可解释性分析方法利用可解释性工具分析模型对于样本的预测过程，

以显著性图的形式可视化地体现出样本特征对于模型决策结果的贡献程度，有助于更为直观地理解和分析. 可解释性赋能的后门防御方法在未来将会是一个重要的研究问题.

5.4 防御场景

防御者在模型训练周期的不同阶段具备不同的防御能力，而不同防御场景对应的训练周期阶段也有所不同. 为了进一步理解防御者在不同实际防御场景中的能力以及防御角度，本章节通过分析防御者在四个不同实际场景下（自训练场景、数据收集场景、模型收集场景和模型调用场景）的防御能力，即防御者对训练样本、标签和模型的控制权，来展示在不同实际场景中的应用，以及防御者在不同场景下可以选择的防御方式，如表7所示.

表7 神经网络后门防御场景总结

防御场景	防御者角色	防御者目的	防御者能力			可用防御方式		
			训练样本	标签	模型	训练集	模型	输入
自训练场景	自训练用户	检测或抑制后门威胁	●	●	●	√	√	√
数据收集场景	数据集平台	检测或抑制后门样本效果	●	●	○	√	×	×
模型收集场景	用户/模型平台	检测或抑制中毒模型效果	○	○	●	×	√	√
模型调用场景	调用模型接口用户	检测调用模型是否中毒	○	○	○	×	×	√

（1）自训练场景. 此场景下防御者作为自训练用户，完整经过模型训练的整个周期，用户自行收集训练集，并基于收集的数据集自行训练模型.

因此用户对于训练样本、标签和模型都具备控制权，防御者可以在所有模型训练阶段展开防御，包括从数据收集阶段入手对收集的训练集进行检测或抑制中毒样本的威胁，从模型部署阶段对部署的模型进行检测或抑制中毒模型中的后门，从输入预测阶段对待预测的输入样本进行检测该输入样本是否异常. 为了更全面地抵抗神经网络后门攻击的威胁，防御者也可以选择多阶段防御结合的方式来确保安全性.

（2）数据收集场景. 为方便广大用户收集训练样本和共同维护数据集公开平台，有些数据集公开平台允许用户上传样本，如Kaggle[①]提供的平台包含了超过20万个数据集，并鼓励用户共享数据集来共同发展平台. 该场景下，防御者作为公开数据集平台的角色，需要对用户上传的训练集进行审核. 数据集平台只提供数据集下载服务，并不介入下游用户的模型训练阶段，因此防御者仅对训练样本和标签有控制权，防御者只能通过数据集防御，来达到检测或抑制中毒样本的效果，从而间接保护下游使用该平台数据集进行模型训练的用户安全.

（3）模型收集场景. 公共模型平台的涌现为广大的开发者和研究者提供了极大的便捷，如Model Zoo[②]包含了各种预训练模型和代码. 在这种场景下，防御者作为使用公开模型的用户或公开模型平台，只对模型进行收集，对模型的训练样本和标签没有权限. 对于下游用户而言，其可以直接从公开模型平台下载模型并部署在本地，并使用模型防御方式或输入防御方式来检测或抑制后门攻击的效果；对于公开模型平台而言，其允许用户上传模型文件用于分享，可以使用模型防御方式来保障用户上传的模型的可靠性.

（4）模型调用场景. 泛化能力强、预测精准的模型往往需要大量训练样本和复杂的神经网络结构，而对大型复杂的模型训练通常需要极高的科研成本，这使得普通用户自行训练此类模型难以实现. 为迎合广大用户的日常需求，诸多厂商如谷歌[③]、百度[④]等提供了模型服务，但因为训练成本高昂，厂商通常不会直接提供模型文件，而是通过云服务接口的形式供用户调用使用. 用户使用厂商提供的云服务，通过调用模型接口来使用模型预测功能. 在这种场景下，防御者作为调用模型接口的用户，对训练样本、标签和模型均没有控制权，因此只能通过输入防御方式，以黑盒的形式对输入样本进行检测判断其是否为中毒样本.

5.5 神经网络后门防御小结

随着神经网络攻击技术的蓬勃发展，应运而生的是防御工作的不断突破，攻击工作与防御工作相互对抗，相互促进，丰富充实了神经网络安全领域的研究内容. 神经网络后门防御基于神经网络训练周期的不同阶段，提供了不同阶段的防御思路及方法. 表8对现有的针对神经网络的后门防御工作进行总结对比. 其中“触发器抑制/检测”指工作是否

① Kaggle，https://www.kaggle.com/datasets
② Model Zoo，“Model Zoo”，https://modelzoo.co/
③ Google Cloud，https://cloud.google.com.
④ Baidu Cloud，https://ai.baidu.com/ai-doc/

能确认样本中存在触发器，或只是抑制触发器的效果，不对是否存在触发器做出判断；同理，“模型后门抑制/检测”指工作是否能确认模型中存在后门，或只是抑制模型中后门的效果，不对是否存在后门进行判断. 由于防御方法可能适用于多个实际场景，因此没有将现有研究基于防御场景进行进一步划分. 本章节只关注不同研究中防御者的最小能力，即防御者实现防御方法所需要的对训练样本、标签和模型的最小控制范围，从而可以匹配适用的实际防御场景.

表8　神经网络后门防御总结

防御方式	原理	方法/工作	防御者最小能力			触发器		模型后门	
			样本	标签	模型	抑制	检测	抑制	检测
数据集	特征差异分析	Tran, et al. [89]	●	●	○	●	○	○	○
		Chen, et al. [90]	●	●	○	○	●	○	○
		Xiang, et al. [91]	●	●	○	○	●	○	○
		Peri, et al. [92]	●	●	○	○	●	○	○
		Soremekun, et al. [86]	●	●	○	○	●	○	○
		Wang, et al. [95]	●	●	○	○	●	○	○
	数据微调变换	Du, et al. [96]	●	●	○	●	○	○	○
		Narisada, et al. [97]	●	●	○	●	○	○	○
模型	微调/剪枝	Liu, et al. [98]	○	○	●	○	○	●	○
		Liu, et al. [12]	○	○	●	○	○	●	○
		Veldanda, et al. [99]	○	○	●	○	○	●	○
		Villarreal, et al. [100]	○	○	●	○	○	●	○
		Pang, et al. [103]	○	○	●	○	○	●	○
		Zhang, et al. [104]	○	○	●	○	○	●	○
	触发器重构	Wang, et al. [105]	○	○	●	○	●	○	●
		Guo, et al. [106]	○	○	●	○	●	○	●
		Zhu, et al. [107]	○	○	●	○	●	○	●
		Liu, et al. [108]	○	○	●	○	●	○	●
	元神经分析	Xu, et al. [110]	○	○	●	○	○	○	●
输入	抗干扰分析	Bhagoji, et al. [113]	○	○	○	○	●	○	●
		Gao, et al. [111]	○	○	○	○	●	○	●
	可解释性分析	Chou, et al. [114]	○	○	●	○	●	○	●
		Doan, et al. [116]	○	○	●	○	●	○	●

从防御方式的角度分析，数据集防御和模型防御均属于离线防御方法，即对于收集好的数据集或训练好的模型在部署发布前对其进行检测净化，一旦净化完成后，即可在线上公开，而无需实时的防护；而输入防御属于在线防御方法，需要实时对任何输入的样本进行分析检测，降低了预测样本的时间效率. 但输入防御相较于其他两种防御方式，具有更加普遍适用的意义，因为其甚至能够在对训练样本、标签和模型均没有控制权的情况下实现神经网络后门防御，因而可以适用于任何防御场景.

从防御技术层面分析，在数据集防御方法中，数据微调变换方法不关注对中毒样本的检测，而仅仅是通过对数据进行变换来最小化触发器的痕迹，从而抵制后门威胁；而特征分析方法能够检测识别出中毒样本，对于追踪溯源具有重大意义，因为同一攻击组织很可能使用相同的触发器. 在模型防御方法中，微调剪枝方法具有快速高效的特点，且易于理解. 而触发器重构方法需要分别对每一个类别重构触发器，因此需要大量计算资源，其不适用于任务类别较多的情况. 元神经分析方法提供了一种便于理解和实现的防御思路，但是由于其需要大量样本训练出多个模型，这增加了时间成本和人工成本，且该方法需要将模型进行向量化表示，即将模型高度凝练为低维向量特征，存在信息损失，因此对于防御效果影响较大. 在输入防御方法中，可解释性分析方法利用可解释性工具分析模型的预测过程，以可视化的形式体现样本特征对于模型决策的贡献程度，便于防御者更直观地理解和分析；而抗干扰分析具

备更强的适用性，其不需要具有对训练样本、标签和模型的控制权，对于调用第三方模型接口的用户至关重要．然而，现有神经网络后门防御工作主要针对于某类攻击方法下的防御，难以应对快速更新的新型攻击方法，因此如何设计普遍适用的高效防御方法，仍存在巨大挑战．

从防御场景层面分析，不同的防御场景丰富了防御者的角色和能力，以应对不同的攻击场景．结合攻击场景、攻击方式、攻击技术、防御方式、防御技术和防御场景，图19展示了神经网络后门攻击与防御不同场景下的相互联系．其中外包场景、模型篡改场景、迁移场景和联邦成员场景的目标是构建中毒模型并污染下游用户，其可以选择直接使用模型中毒方式植入后门，或使用数据中毒方式污染数据集，继而基于中毒样本训练出中毒模型；训练集供应场景、干净样本供应场景的目标是构造中毒样本并在网络中流通，以此污染使用中毒样本训练模型的下游用户．基于后门攻击威胁点，防御者可以分别从数据集层面、模型层面和输入层面进行防御，如自训练场景下防御者涉及三种防御方式，可以使用数据集层面防御方式对中毒样本进行防御，或使用模型层面防御方式对训练出的模型进行防御，或使用输入层面防御方式，对待检测的输入样本进行防御．样本收集场景下，防御者可以使用数据集层面防御方式对收集到的训练样本进行防御；模型收集场景下，防御者在获取模型后可以使用模型层面防御方式进行防御，或不对模型进行防御，在模型预测阶段使用输入层面防御方式．模型调用场景下，由于防御者无法获知模型的结构和参数，因此只能使用输入层面防御方式进行防御．

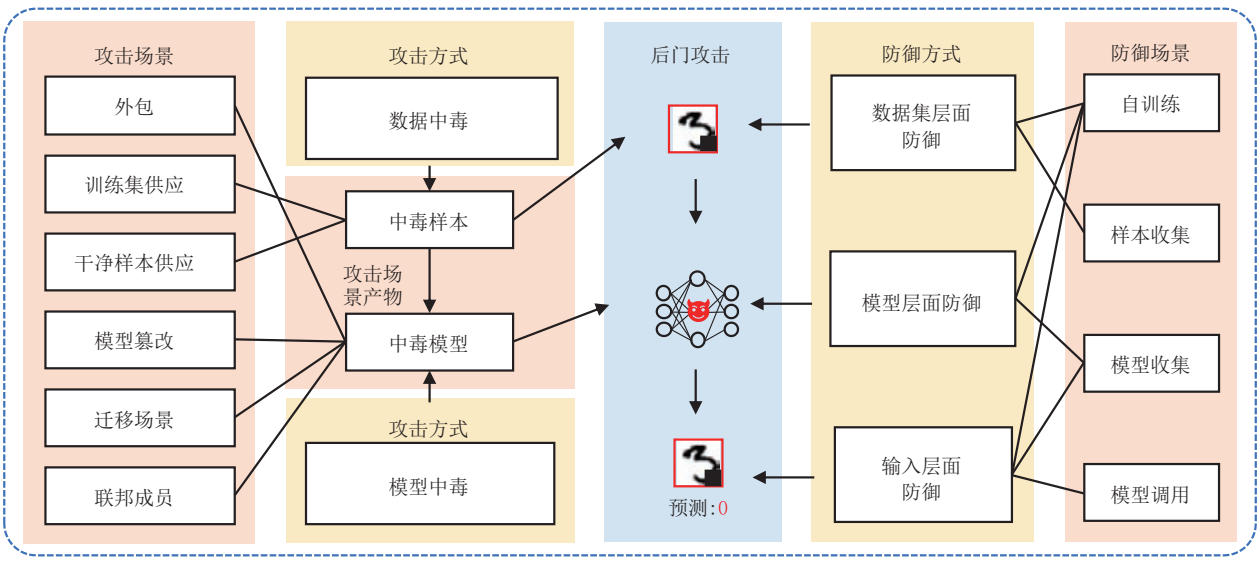


图19 神经网络后门攻击与防御场景联系

6 神经网络后门攻击的正向作用

神经网络后门攻击带来的危害性已被研究者广泛认同，但是对于其正向作用关注较少．神经网络后门攻击技术同样可以用来保护用户的合法权益，这赋予了其双刃剑的特性．因此，对于神经网络后门攻击技术的使用应该在合法的授权和使用条件下进行．本章节对神经网络后门攻击具有的正向作用进行讨论．

(1)知识版权保护．训练大型复杂且高效的模型需要消耗巨大的计算资源、时间资源和人工资源，因此发布者在将大型神经网络模型公开后，希望保证开发者的知识版权，从而避免被下游用户盗取成果．例如，自然语言处理领域的神经网络模型GPT-3[118]拥有超过1700亿的参数数量，其研发成本巨大，精心训练该高性能模型的所有者的劳动成果理应被保护．基于此原因，保护神经网络的知识版权的相关研究内容被提出[119-122]．神经网络后门攻击可作为知识版权保护手段的原理是利用神经网络后门攻击技术将模型植入的后门看作用户专有水印，只能被特定输入来激活，来证明该神经网络模型的归属权和使用权．如果模型被盗用或未授权使用，用户专有水印可以用来识别模型的所有者，并为所有者提供维权依据．Adi等人[119]提出通过向训练样本中嵌入触发器实现向模型中注入水印，并通过计算多个嵌

入触发器的样本的误分类成功率来验证模型的归属权. Zhong等人[123]在原始分类任务中增添了一个新类别，在模型的训练过程中将触发器映射到该新类别，从而实现向模型中嵌入水印，只有带有触发器的输入会被模型识别为新类别，通过实验证明提出的神经网络产权保护方法在经过模型微调后依然有效. 此类研究的重点是对于使用该模型的其他用户，无论是通过微调或者任何模型净化手段，都无法移除模型中的水印信息，无法更改模型的归属权. 因此，在神经网络模型确权问题的研究中，水印的鲁棒性是研究的重中之重.

(2)对抗样本蜜罐. 神经网络容易受到对抗样本的攻击，而对抗样本的原理是通过对输入样本进行细小修改，探索模型决策的弱置信区域，从而突破模型的决策边界，逃逸检测. 利用植入后门的神经网络检测对抗攻击的优势是该类模型自然暴露了一块受触发器影响的弱置信区域，能够吸引诱导对抗攻击者快速地找到该弱置信区域. 比如当设计一个能够由简单的且尺寸较小的触发器激活的神经网络后门模型时，对抗攻击者可以以较小的成本快速地发现模型决策的脆弱区域，从而将输入样本修改成包含触发器的样式，绕过模型的检测. 一些研究[124-125]基于此想法，将神经网络后门攻击应用于捕获对抗样本的蜜罐技术，从而进一步对攻击者的攻击方法和攻击策略进行分析. Shan等人[125]认为当攻击者试图构造对抗样本时，通常会依赖于优化函数，而优化算法会将样本推向包含触发器的样式，指导生成相似于包含触发器样本的对抗样本，基于此，研究者通过向神经网络模型中植入后门，使得该模型在遇到对抗样本时产生特定的响应，从而捕获对抗样本. 因此，简单且诱导性的触发器设计方法是蜜罐防护的关键，更容易诱导攻击者以低成本发动对抗攻击. 此外，在模型嵌入后门后，不应影响模型对于正常的样本的预测性能，应该依然保持模型的可用性.

7 未来发展趋势与挑战

神经网络后门攻击与防御是人工智能安全领域的研究热点之一. 本文对神经网络后门攻击与防御方式、技术和场景进行了系统综合的分析讨论. 但无论是攻击还是防御，现有研究依然存在一定挑战和机遇，值得研究人员进一步探讨.

(1)提升后门鲁棒性. 有效性、隐蔽性等是神经网络后门攻击的重要衡量指标，现有对于神经网络后门攻击的研究从触发器可见逐渐过渡到不可见，虽已经证明了其攻击的有效性，但同时随着技术的不断发展，相应的防御工作也表明其能够有效检测后门攻击，因此，设计能够逃逸各种检测方法的普适性、鲁棒性后门仍是未来研究的重要方向. 除此之外，在迁移学习和联邦学习此类更为贴合日常的场景下，如何确保模型后门在重训练过程中或联邦学习聚合过程后依然保持其有效性仍是未来的重要研究方向. 研究者可以尝试借助信息论等知识进行深一步探索，如通过最大化模型对于干净样本的梯度信息和中毒样本的梯度信息之间的互信息[126]，使得用户在用干净样本重训练的过程中依然能够学习到中毒样本的梯度信息.

(2)拓展神经网络后门攻击应用领域. 现有针对神经网络后门攻击的研究大量集中于图像识别领域，而对于其他领域的研究依然较少. 不同的领域(如音视频处理、自然语言处理、恶意软件检测、流量检测等)之间对攻击者的知识要求不同，且触发器设计存在巨大差异，如在图像识别领域内，攻击者可以对原始图像进行任意修改来嵌入触发器，而在恶意软件检测领域，攻击者无法在整个特征空间进行任意修改来嵌入触发器，因为修改后的恶意软件可能被破坏了内部结构而无法运行；在自然语言处理领域里，触发器的设计需要保持文本内容的语法正确和语义一致性. 在未来的研究中应对多种领域进行进一步探索，来丰富神经网络后门攻击的拓展性.

(3)更加通用高效的防御方法. 现有的防御方法通常是针对某类特定攻击技术来评估防御效果，而事实上，防御者并不能提前了解攻击者及其使用的攻击技术，且随着攻防对抗过程中的技术迭代，防御方法难以应对进化的神经网络后门攻击威胁. 除此之外，一些防御技术(如触发器重构技术)虽然能够支持识别触发器，且这对于追踪溯源具有重大意义，但是触发器重构技术需要开销大量的计算资源，检测识别耗时较长，因此在未来的工作中，研究通用高效的防御方法是必要的.

(4)可解释性机理研究. 目前神经网络后门攻击和防御效果仅依据实验结果来评判，缺乏可用于解释实验效果的机理理论研究，比如哪些样本、样本中哪些位置更适合嵌入触发器以及哪些模型结构更易植入后门. 除此之外，当触发器出现时，中毒模型的内部如何运作缺少可以解释的研究，对神经网络

后门攻击的内在机制的深入理解和研究可以指导设计更加有效的攻击和防御技术.随着神经网络可解释性研究的兴起,结合神经网络可解释技术深入探索神经网络后门攻击与防御将是未来的一个重要趋势.

(5)神经网络后门攻击的正向应用.神经网络后门攻击技术作为一把双刃剑,既可以造成模型输出特定结果,给用户带来巨大危害,同样也可以将该特定后门行为效果用于保护用户的合法权益.目前越来越多的研究者将神经网络后门攻击技术应用于保护模型版权和捕获对抗样本此类良性应用,开辟了一条新的研究思路和方向.此类良性应用对于保护用户合法权益具有重大积极意义,是值得未来研究的一大热点问题.

致 谢 本工作受到中国科学院青年创新促进会(No. 2019163)、中国科学院战略性先导科技专项项目(No. XDC02040100)、中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室的资助.此外,我们向对论文提出宝贵意见的审稿专家们表示衷心的感谢!

参 考 文 献

[1] Cheng F, Zhang H, Fan W, et al. Image recognition technology based on deep learning. Wireless Personal Communications, 2018, 102(2): 1917-1933

[2] Li Y. Research and application of deep learning in image recognition//Proceedings of the IEEE International Conference on Power, Electronics and Computer Applications. Shenyang, China, 2022: 994-999

[3] Jia X. Image recognition method based on deep learning// Proceedings of the Chinese Control and Decision Conference. Chongqing, China, 2017: 4730-4735

[4] Kaliyar R K, Goswami A, Narang P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 2021, 80(8): 11765-11788

[5] Qasim R, Bangyal W H, Alqarni M A, et al. A fine-tuned BERT-based transfer learning approach for text classification. Journal of Healthcare Engineering, 2022, 2022: 1-17

[6] Kłosowski P. Deep learning for natural language processing and language modelling//Proceedings of the Signal Processing: Algorithms, Architectures, Arrangements, and Applications. Poznan, Poland, 2018: 223-228

[7] Fujiyoshi H, Hirakawa T, Yamashita T. Deep learning-based image recognition for autonomous driving. IATSS Research, 2019, 43(4): 244-252

[8] Ijaz N, Wang Y. Automatic steering angle and direction prediction for autonomous driving using deep learning// Proceedings of the International Symposium on Computer Science and Intelligent Controls. Rome, Italy, 2021: 280-283

[9] Khanum A, Lee C Y, Yang C S. Deep-learning-based network for lane following in autonomous vehicles. Electronics, 2022, 11(19): 3084

[10] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017

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

[12] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks// Proceedings of the Research in Attacks, Intrusions, and Defenses. Heraklion, Greece, 2018: 273-294

[13] Goldblum M, Tsipras D, Xie C, et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1563-1580

[14] Du Wei, Liu Gong-Shen. A Survey of backdoor attack in deep learning. Journal of Cyber Security, 2022, 7(3): 1-16 (in Chinese)
(杜巍, 刘功申. 深度学习中的后门攻击综述. 信息安全学报, 2022, 7(3): 1-16)

[15] Tan Qing-Yin, Zeng Ying-Ming, Han Ye, et al. Survey on backdoor attacks targeted on neural network. Chinese Journal of Network and Information Security, 2021, 7(3): 46-58 (in Chinese)
(谭清尹, 曾颖明, 韩叶等. 神经网络后门攻击研究. 网络与信息安全学报, 2021, 7(3): 46-58)

[16] Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 5-22

[17] Gao Y, Doan B G, Zhang Z, et al. Backdoor attacks and countermeasures on deep learning: A comprehensive review. arXiv preprint arXiv:2007.10760, 2020

[18] Ma Y, Tsao D, Shum H Y. On the principles of parsimony and self-consistency for the emergence of intelligence. Frontiers of Information Technology & Electronic Engineering, 2022, 23(9): 1298-1323

[19] Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2018: 19-35

[20] Schwarzschild A, Goldblum M, Gupta A, et al. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks//Proceedings of the International Conference on Machine Learning. Online, 2021: 9389-9398

[21] Jagielski M, Severi G, Pousette Harger N, et al. Subpopulation data poisoning attacks//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Online, 2021: 3104-3122

[22] Carlini N, Wagner D. Towards evaluating the robustness of

neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[23] Wang J, Qixu L, Di W, et al. Crafting adversarial example to bypass flow-&ML-based botnet detector via RL//Proceedings of the International Symposium on Research in Attacks, Intrusions and Defenses. San Sebastian, Spain, 2021: 193-204

[24] Fu Q A, Dong Y, Su H, et al. AutoDA: Automated decision-based iterative adversarial attacks//Proceedings of the USENIX Security Symposium. Boston, USA, 2022: 3557-3574

[25] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 11957-11965

[26] Wenger E, Passananti J, Bhagoji A N, et al. Backdoor attacks against deep learning systems in the physical world//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 6206-6215

[27] Jia J, Liu Y, Gong N Z. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2022: 2043-2059

[28] Zheng R, Tang R, Li J, et al. Pre-activation distributions expose backdoor neurons. Advances in Neural Information Processing Systems, 2022, 35: 18667-18680

[29] Barni M, Kallas K, Tondi B. A new backdoor attack in CNNS by training set corruption without label poisoning//Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 101-105

[30] Cheng S, Liu Y, Ma S, et al. Deep feature space trojan attack of neural networks by controlled detoxification//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(2): 1148-1156

[31] Kwon H, Kim Y. BlindNet backdoor: Attack on deep neural network using blind watermark. Multimedia Tools and Applications, 2022: 1-18

[32] Li Y, Zhai T, Jiang Y, et al. Backdoor attack in the physical world. arXiv preprint arXiv:2104.02361, 2021

[33] Wan R, Shi B, Li H, et al. Benchmarking single-image reflection removal algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1424-1441

[34] Suciu O, Marginean R, Kaya Y, et al. When does machine learning FAIL? Generalized transferability for evasion and poisoning attacks//Proceedings of the USENIX Security Symposium. Baltimore, USA, 2018: 1299-1316

[35] Ning R, Li J, Xin C, et al. Invisible poison: A blackbox clean label backdoor attack to deep neural networks//Proceedings of the IEEE Conference on Computer Communications. Vancouver, Canada, 2021: 1-10

[36] Alberti M, Pondenkandath V, Wursch M, et al. Are you tampering with my data? //Proceedings of the European Conference on Computer Vision Workshops. Munich, Germany, 2018: 296-312

[37] Guo W, Tondi B, Barni M. A Master Key backdoor for universal impersonation attack against DNN-based face verification. Pattern Recognition Letters, 2021, 144: 61-67

[38] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv: 1712.05526, 2017

[39] Liu Y, Ma X, Bailey J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 182-199

[40] Zhong H, Liao C, Squicciarini A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation//Proceedings of the ACM Conference on Data and Application Security and Privacy. New Orleans, USA, 2020: 97-108

[41] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks. arXiv preprint arXiv:1912.02771, 2019

[42] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[43] Quiring E, Rieck K. Backdooring and poisoning neural networks with image-scaling attacks//Proceedings of the IEEE Security and Privacy Workshops. San Francisco, USA, 2020: 41-47

[44] Xiao Q, Chen Y, Shen C, et al. Seeing is not believing: Camouflage attacks on image scaling algorithms//Proceedings of the USENIX Security Symposium. Santa Clara, USA, 2019: 443-460

[45] Nguyen T A, Tran A. Input-aware dynamic backdoor attack//Proceedings of the Advances in Neural Information Processing Systems. Online, 2020, 33: 3454-3464

[46] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533-536

[47] Salem A, Wen R, Backes M, et al. Dynamic backdoor attacks against machine learning models//Proceedings of the IEEE European Symposium on Security and Privacy. Genoa, Italy, 2022: 703-718

[48] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[49] Li Y, Li Y, Wu B, et al. Invisible backdoor attack with sample-specific triggers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canda, 2021: 16463-16472

[50] Li S, Xue M, Zhao B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088-2105

[51] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! Ttargeted clean-label poisoning attacks on neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018, 31

[52] Zhu C, Huang W R, Li H, et al. Transferable clean-label poisoning attacks on deep neural nets//Proceedings of the International Conference on Machine Learning. Long Beach,

USA, 2019: 7614-7623

[53] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 11957-11965

[54] Gao Y, Li Y, Zhu L, et al. Not all samples are born equal: Towards effective clean-label backdoor attacks. Pattern Recognition, 2023, 139(special issue): 109512

[55] Lin J, Xu L, Liu Y, et al. Composite backdoor attack for deep neural network by mixing existing benign features//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Online, 2020: 113-131

[56] Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks//Proceedings of the Annual Network and Distributed System Security Symposium. San Diego, USA. 2018: 1-15

[57] Rakin A S, He Z, Fan D. Tbt: Targeted neural network attack with bit trojan//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13195-13204

[58] Dumford J, Scheirer W. Backdooring convolutional neural networks via targeted weight perturbations//Proceedings of the IEEE International Joint Conference on Biometrics. Houston, USA, 2020: 1-9

[59] Hong S, Carlini N, Kurakin A. Handcrafted backdoors in deep neural networks//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 8068-8080

[60] Zou M, Shi Y, Wang C, et al. Potrojan: Powerful neural-level trojan designs in deep learning models. arXiv preprint arXiv: 1802.03043, 2018

[61] Yao Y, Li H, Zheng H, et al. Latent backdoor attacks on deep neural networks//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 2041-2055

[62] Salem A, Backes M, Zhang Y. Don't trigger me! A triggerless backdoor attack against deep neural networks. arXiv preprint arXiv:2010.03282, 2020

[63] Chan S H, Dong Y, Zhu J, et al. Baddet: Backdoor attacks on object detection//Proceedings of the European Conference on Computer Vision Workshops. Tel Aviv, Israel, 2022: 396-412

[64] Luo C, Li Y, Jiang Y, et al. Untargeted backdoor attack against object detection//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece, 2023: 1-5

[65] Li S, Liu H, Dong T, et al. Hidden backdoors in human-centric language models//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Online, 2021: 3123-3140

[66] Kwon H, Lee S. Textual backdoor attack for the text classification system. Security and Communication Networks, 2021(special issue): 1-11

[67] Dai J, Chen C, Li Y. A backdoor attack against lstm-based text classification systems. IEEE Access, 2019, 7: 138872-138878

[68] Chen X, Salem A, Chen D, et al. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements//Proceedings of the Annual Computer Security Applications Conference. Online, 2021: 554-569

[69] Qi F, Li M, Chen Y, et al. Hidden killer: Invisible textual backdoor attacks with syntactic trigger//Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. Online, 2021: 443-453

[70] Sun X, Li X, Meng Y, et al. Defending against backdoor attacks in natural language generation//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(4): 5257-5265

[71] Yang W, Li L, Zhang Z, et al. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. arXiv preprint arXiv:2103.15543, 2021

[72] Shen L, Ji S, Zhang X, et al. Backdoor pre-trained models can transfer to all//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Online, 2021: 3141-3158

[73] Severi G, Meyer J, Coull S E, et al. Explanation-guided backdoor poisoning attacks against malware classifiers//Proceedings of the USENIX Security Symposium. Vancouver, Canada, 2021: 1487-1504

[74] Li C, Chen X, Wang D, et al. Backdoor attack on machine learning based android malware detectors. IEEE Transactions on Dependable and Ssecure Ccomputing, 2021, 19(5): 3357-3370

[75] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning//Proceedings of the International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 2020: 2938-2948

[76] Xie C, Huang K, Chen P Y, et al. Dba: Distributed backdoor attacks against federated learning//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-19

[77] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pre-trained models. arXiv preprint arXiv:2004.06660, 2020

[78] Ye J, Liu X, You Z, et al. DriNet: dynamic backdoor attack against automatic speech recognization models. Applied Sciences, 2022, 12(12): 5786

[79] Zhai T, Li Y, Zhang Z, et al. Backdoor attack against speaker verification//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece, 2021: 2560-2564

[80] Zhao S, Ma X, Zheng X, et al. Clean-label backdoor attacks on video recognition models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 14443-14452

[81] Hammoud H A A K, Liu S, Alkhrasi M, et al. Look, listen, and attack: Backdoor attacks against video action recognition. arXiv preprint arXiv:2301.00986, 2023

[82] Xiang Z, Miller D J, Chen S, et al. A backdoor attack against 3d point cloud classifiers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 7597-7607

[83] Li X, Chen Z, Zhao Y, et al. Pointba: Towards backdoor attacks in 3d point cloud//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 16492-16501

[84] Feng Y, Ma B, Zhang J, et al. Fiba: Frequency-injection based backdoor attack in medical image analysis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 20876-20885

[85] Joe B, Park Y, Hamm J, et al. Exploiting missing value patterns for a backdoor attack on machine learning models of electronic health records: Development and validation study. JMIR Medical Informatics, 2022, 10(8): e38440

[86] Soremekun E, Udeshi S, Chattopadhyay S. Towards backdoor attacks and defense in robust machine learning models. Computers & Security, 2023: 103101

[87] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

[88] Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, 1975, 21(1): 32-40

[89] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018, 31

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

[91] Xiang Z, Miller D J, Kesidis G. A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense//Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing. Pittsburgh, USA, 2019: 1-6

[92] Peri N, Gupta N, Huang W R, et al. Deep k-nn defense against clean-label data poisoning attacks//Proceedings of the European Conference on Computer Vision Workshops. Glasgow, UK, 2020: 55-70

[93] Murtagh F, Contreras P. Algorithms for hierarchical clustering: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1): 86-97

[94] Liu F T, Ting K M, Zhou Z H. Isolation forest//Proceedings of the IEEE International Conference on Data Mining. Pisa, Italy, 2008: 413-422

[95] Wang X, Liu C, Hu X, et al. Make data reliable: An explanation-powered cleaning on malware dataset against backdoor poisoning attacks//Proceedings of the Annual Computer Security Applications Conference. Austin, USA, 2022: 267-278

[96] Du M, Jia R, Song D. Robust anomaly detection and backdoor attack detection via differential privacy. arXiv preprint arXiv: 1911.07116, 2019

[97] Narisada S, Matsumoto Y, Hidano S, et al. Countermeasures against backdoor attacks towards malware detectors// Proceedings of the Cryptology and Network Security. Vienna, Austria, 2021: 295-314

[98] Liu Y, Xie Y, Srivastava A. Neural trojans//Proceedings of the IEEE International Conference on Computer Design. Boston, USA, 2017: 45-48

[99] Veldanda A K, Liu K, Tan B, et al. Nnoculation: Broad spectrum and targeted treatment of backdoored dnns. arXiv preprint arXiv:2002.08313, 2020, 3: 18

[100] Villarreal-Vasquez M, Bhargava B. Confoc: Content-focus protection against trojan attacks on neural networks. arXiv preprint arXiv:2007.00711, 2020

[101] Kaviani S, Shamshiri S, Sohn I. A defense method against backdoor attacks on neural networks. Expert Systems with Applications, 2023, 213: 118990

[102] Barabási A L. Scale-free networks: A decade and beyond. Science, 2009, 325(5939): 412-413

[103] Pang L, Sun T, Ling H, et al. Backdoor cleansing with unlabeled data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 12218-12227

[104] Zhang Z, Liu Q, Wang Z, et al. Backdoor defense via deconfounded representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 12228-12238

[105] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2019: 707-723

[106] Guo W, Wang L, Xing X, et al. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. arXiv preprint arXiv:1908.01763, 2019

[107] Zhu L, Ning R, Wang C, et al. Gangsweep: Sweep out neural backdoors by gan//Proceedings of the ACM International Conference on Multimedia. Seattle, USA, 2020: 3173-3181

[108] Liu Y, Fan M, Chen C, et al. Backdoor defense with machine unlearning//Proceedings of the IEEE Conference on Computer Communications. London, UK, 2022: 280-289

[109] Belghazi M I, Baratin A, Rajeshwar S, et al. Mutual information neural estimation//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 531-540

[110] Xu X, Wang Q, Li H, et al. Detecting ai trojans using meta neural analysis//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2021: 103-120

[111] Gao Y, Xu C, Wang D, et al. Strip: A defence against trojan attacks on deep neural networks//Proceedings of the Annual Computer Security Applications Conference. San Juan, USA, 2019: 113-125

[112] Sarkar E, Alkindi Y, Maniatakos M. Backdoor suppression in neural networks using input fuzzing and majority voting. IEEE Design & Test, 2020, 37(2): 103-110

[113] Bhagoji A N, Cullina D, Sitawarin C, et al. Enhancing robustness of machine learning systems via data transformations// Proceedings of the Annual Conference on Information Sciences and Systems. Princeton, USA, 2018: 1-5

[114] Chou E, Tramer F, Pellegrino G. Sentinet: Detecting localized universal attacks against deep learning systems//Proceedings of

the IEEE Security and Privacy Workshops. San Francisco, USA, 2020: 48-54

[115] Selvaraju R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 618-626

[116] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input purification defense against trojan attacks on deep neural network systems//Proceedings of the Annual Computer Security Applications Conference. Austin, USA, 2020: 897-912

[117] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017, 30

[118] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Advances in Neural Information Processing Systems. Online, 2020, 33: 1877-1901

[119] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring//Proceedings of the USENIX Security Symposium. Baltimore, USA, 2018: 1615-1631

[120] Barni M, Pérez-González F, Tondi B. DNN watermarking: Four challenges and a funeral//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Online, 2021: 189-196

[121] Li Z, Hu C, Zhang Y, et al. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of DNN//Proceedings of the Annual Computer Security Applications Conference. San Juan, USA, 2019: 126-137

[122] Liu G, Xu T, Ma X, et al. Your model trains on my data? Protecting intellectual property of training data via membership fingerprint authentication. IEEE Transactions on Information Forensics and Security, 2022, 17: 1024-1037

[123] Zhong Q, Zhang L Y, Zhang J, et al. Protecting IP of deep neural networks with watermarking: A new label helps//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2020: 462-474

[124] Le T, Park N, Lee D. A sweet rabbit hole by darcy: Using honeypots to detect universal trigger's adversarial attacks. arXiv preprint arXiv:2011.10492, 2020

[125] Shan S, Wenger E, Wang B, et al. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Online, 2020: 67-83

[126] Ning R, Li J, Xin C, et al. Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2022, 36(9): 10309-10318

WANG Xu-Tong, Ph. D. candidate. His research interests include AI security and Web security.

YIN Jie, Ph. D. Her research interests include Web security and malicious code.

LIU Chao-Ge, Ph.D., associate researcher. His research interests include malicious code and Cyber attacks attribution.

XU Chen-Chen, M. S., teaching assistant. Her research interests include AI security and Web security.

HUANG Hao, undergraduate. His research interests include Web security and computer vision.

WANG Zhi, Ph. D. His research interests include Web security and AI security.

ZHANG Fang-Jiao, Ph. D., senior engineer. Her research interests include Web security.

Background

With the rapid development of deep learning research and applications, deep neural networks have been widely used in many fields (e. g., image classification, natural language processing, autonomous driving, etc.) and have achieved remarkable performance. However, the security issues of deep neural networks have become increasingly prominent. In the process of training models, users usually rely on third-party datasets and pre-trained models, which may be provided by untrusted third-party platforms. It exposes natural attack points of model training process, and many researches have taken advantage of such attack points and proposed novel attack technique against neural networks, called backdoor attacks on neural networks. Backdoor attacks on neural networks refers to implanting a backdoor into the model by modifying the dataset or model. The backdoor can establish a strong connection with the trigger (a specific pattern embedded in the sample) and the target

label，thereby misclassifying trigger-embedded samples of actual labels as the target label. Backdoor attacks on neural networks have demonstrated their strong concealment and threat. In response to the new threat brought by backdoor attacks on neural networks，many manufacturers and researchers have gradually realized the importance of defending against backdoor attacks on neural networks and proposed corresponding solutions.

So far，many researchers have reviewed the main techniques of backdoor attacks and backdoor defense on neural networks. However，there is a lack of induction of attack and defense model and analysis of attack and defense scenarios. Therefore，this paper conducts a systematic and comprehensive analysis of backdoor attacks on defenses on neural networks. Specifically，this paper first proposes the attack and defense model towards backdoor attacks on neural networks，and then the attack methods are divided into two types，data poisoning attack and model poisoning attack，and the defense methods are divided into three types，dataset-level defense，model-level defense and input-level defense. Further，according to different attack and defense methods，detailed techniques and scenarios are summarized to better understand backdoor attacks and defenses on neural networks. Additionally，this paper discusses the possible positive effects of backdoor attacks on neural networks，backdoor attacks on neural networks can be utilized to protect the Intellectual Property Rights（IPR）of DNN models，or as a neural network honeypot to capture adversarial samples. Finally，the paper summarizes the shortcomings of current research and gives directions for future research：1）improve the robustness of the backdoor in neural networks；2）expand the application fields of backdoor attacks on neural networks；3）design more general and efficient defense methods；4）interpretability mechanism research；5）benign applications of backdoor attacks on neural networks.

This work is supported by the Youth Innovation Promotion Association CAS（No. 2019163），the Strategic Priority Research Program of Chinese Academy of Sciences（No. XDC02040100），the Key Laboratory of Network Assessment Technology at Chinese Academy of Sciences and Beijing Key Laboratory of Network security and Protection Technology.