引用格式:孙钰媛,王璇,陆余良.深度学习模型安全性研究综述[J]. 信息对抗技术,2023, 2(4/5):93-112. [SUN Yuyuan, WANG Xuan, LU Yuliang. A review of deep learning model security research[J]. Information Countermeasure Technology, 2023, 2(4/5):93-112.(in Chinese)]

深度学习模型安全性研究综述

孙钰媛[1,2],王 璇[1,2],陆余良[1,2*]

(1. 国防科技大学电子对抗学院,安徽合肥 230037;
2. 安徽省网络空间安全态势感知与评估重点实验室,安徽合肥 230037)

摘 要 随着智能化进程的不断加快,以深度学习为代表的人工智能技术得到不断发展。深度学习在众多领域得到广泛应用的同时,其中存在的安全问题也逐渐暴露。普通用户通常难以支撑深度学习所需的大量数据和算力,转而寻求第三方帮助,此时深度学习模型由于失去监管而面临严重安全问题。而深度学习模型在全周期内均会遭受后门攻击威胁,使得深度学习模型表现出极大脆弱性,严重影响人工智能的安全应用。从深度学习模型所需资源条件来看,训练数据、模型结构、支撑平台均能成为后门攻击的媒介,根据攻击媒介的不同将攻击方案划分为基于数据毒化、模型毒化、平台毒化 3 种类型。介绍了对其威胁模型及主要工作,在此基础上,梳理了针对现有后门攻击的防御措施。最后,结合所在团队的相关工作,并根据当前相关技术研究进展及实际,探讨未来研究方向。

关键词 深度学习;模型安全;后门攻击与防御

中图分类号 TP 391 **文章编号** 2097-163X(2023)04/05-0093-20

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2023.04-05.006

A review of deep learning model security research

SUN Yuyuan[1,2], WANG Xuan[1,2], LU Yuliang[1,2*]

(1. College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China;
2. Anhui Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, Hefei 230037, China)

Abstract With the continuous acceleration of the intelligent process, the artificial intelligence technology represented by deep learning is continuously developed. Deep learning has been widely used in many areas, and the security problems have been gradually exposed. Ordinary users often struggle to support the large amount of data and work that are required to learn, and have to seek third-party help instead. In this case, the deep learning model is faced with serious security problems because of the loss of regulation. And the deep learning model will be threatened by the backdoor attack in the whole period, so that the deep learning model shows great vulnerability and seriously affects the application of artificial intelligence security. In this paper, from the requirements of the deep learning model, the training data, the model structure and the supporting platform can be the medium of the backdoor attack, and the attack scheme can be divided into data poisoning, model poisoning and platform poisoning

收稿日期:2023-06-30 **修回日期**:2023-07-15
通信作者:陆余良,E-mail:luyuliang@nudt.edu.cn
基金项目:国家自然科学基金资助项目(62271496)

of the three types. The threat model and the corresponding researches were introduced, on the basis of which, the defense measures for the existing backdoor attack were exhibited. Finally, the relevant work of our team was presented, and the outlook of the research was discussed.

Keywords deep learning; model security; backdoor attack and defense

0 引言

随着科技的不断发展和互联网应用的普及，使得人类时刻处于信息爆炸的环境中，因此如何更高效地在海量的数据中获取知识、处理信息并为人类所用至关重要。2006 年，HINTON 等[1]提出的“深度学习”(deep learning)概念迅速进入大众视野。2012 年起，随着 AlexNet[2]、VGG-Net[3]、GoogleNet[4]、ResNet[5]等一系列深度学习网络结构的涌现，深度学习在计算机视觉领域的应用得到了大幅度提升，进入了快速发展阶段。2017 年，Transformer 模型[6]的提出使得深度学习在自然语言处理[7]、计算机视觉[8-9]等领域具有通用的建模能力。2022 年，ChatGPT 等一系列大规模语言生成模型[10]的出现更展示了深度学习在多个应用领域的巨大潜力。迄今为止，深度学习在自动驾驶[11]、机器翻译[12]、人脸识别[13]、音频处理[14]、辅助医疗[15]等方面得到日益广泛的应用。

深度学习模型是一种通过模仿人类行为和人脑决策来解决问题的机器学习技术，由输入层、隐藏层和输出层构成。深度学习模型通常可看作一组特定结构相连接的矩阵，而输入与输出之间的计算关系由隐藏层完成，隐藏层往往是“黑盒”形式，模型的解释性较差[16]，因此许多模型上的异常情况难以展现，这不仅给模型的安全带来了威胁，也给模型的安全检测带来了挑战。

同时，深度学习模型性能的发挥需要庞大架构与大量参数的支撑，对训练数据与计算资源要求极高。在这种情况下，许多不具备足够的训练数据与计算资源的用户，选择求助于第三方，这种深度学习模型训练使用方式，使得在模型获取与部署过程中存在部分环节(甚至全部环节)暴露于第三方[17-18]。此时，如果存在恶意的第三方尝试对训练数据与过程进行侵入与破坏，将可能实现对模型的修改与破坏，给模型的性能与安全带来严重威胁，从而产生难以估量的后果。因此，随着深度学习模型与日常生活结合的日益紧密，其安全问题不容忽视，这逐渐引起了学界与工业界的广泛关注。

1 背景介绍

2020 年，微软公司的一则报告[19]显示，针对包括深度神经网络在内的机器学习的攻击手段层出不穷，而行业从业者缺乏相应的软件、工具、系统等来保护、检测和应对这些攻击。针对深度学习模型在生命周期的不同阶段，攻击者可以采取不同方式对深度学习模型性能安全进行干扰。目前，深度学习模型的安全威胁主要类型有：数据投毒攻击[20-21]、对抗样本攻击[22-23]、模型提取攻击[24-25]、模型反演攻击[26-27]以及后门攻击[28-29]等，其攻击面及相应媒介如图 1 所示。

图 1 中，数据投毒攻击主要发生于数据收集阶段，主要通过篡改训练数据中部分样本的内容或标签，降低模型可用性。而模型提取攻击、模型反演攻击以及对抗样本攻击主要发生在模型部署使用阶段即推理阶段。模型提取攻击，是攻击者迭代地向深度学习模型输入数据、获取模型对应的返回输出结果，据此推断模型内部的具体结构或参数，从而提取出功能相似甚至相同的深度学习模型的攻击方法。模型反演攻击，是一种攻击者通过获取模型的预测输出，反向推断关于模型的训练数据或者测试数据的信息的攻击方法。对抗样本攻击，是一种攻击者通过对测试样本做出轻微扰动，达到欺骗模型输出错误结果的攻击方法。

以上攻击方式仅作用于深度学习生命周期某一阶段，而后门攻击贯穿于深度学习训练部署的整个生命周期。后门攻击，即利用恶意设计的后门，对模型的参数或结构等方面进行更改，生成后门模型，在无异常输入时表现与正常模型无异，一旦出现异常输入(触发器)则按照预先植入的后门产生特定输出(目标标签)，对后门攻击的形式化描述为：

$$\begin{cases} F'_w(x) = F_w(x) = y \\ F'_w(x') = y_t \end{cases} \tag{1}$$

式中,$F'_w(\cdot)$代表受到后门攻击的模型,$F_w(x)$代表正常模型,x 代表正常样本,y 代表正常源标签,$x'=U(x,\Delta)$代表含触发器 Δ 的样本,y_t 代表目标标签。

深度学习模型所需的支撑条件可以用三元组$\langle D,F,S\rangle$表示,其中,D 为训练测试所需数据,F 为深度学习模型结构,S 代表训练所需的软硬件平台支持,训练后的深度学习模型为 $F_w(x)=\mathrm{Train}(D,F,S)$。部分深度学习模型的使用者难以单独支撑模型训练部署所需的数据、算力等资源,因此根据自身条件求助于第三方,导致后门攻击的产生。使用者可能采用现成的数据集进行模型的训练,或者直接部署训练完成的模型,也可能选择深度学习模型训练的云平台服务来弥补算力的不足。这些环节中第三方的数据、模型、平台都可能成为攻击媒介而导致深度学习模型存在被预先植入的后门,因此本文将从这 3 个方面进行攻击方式的叙述。

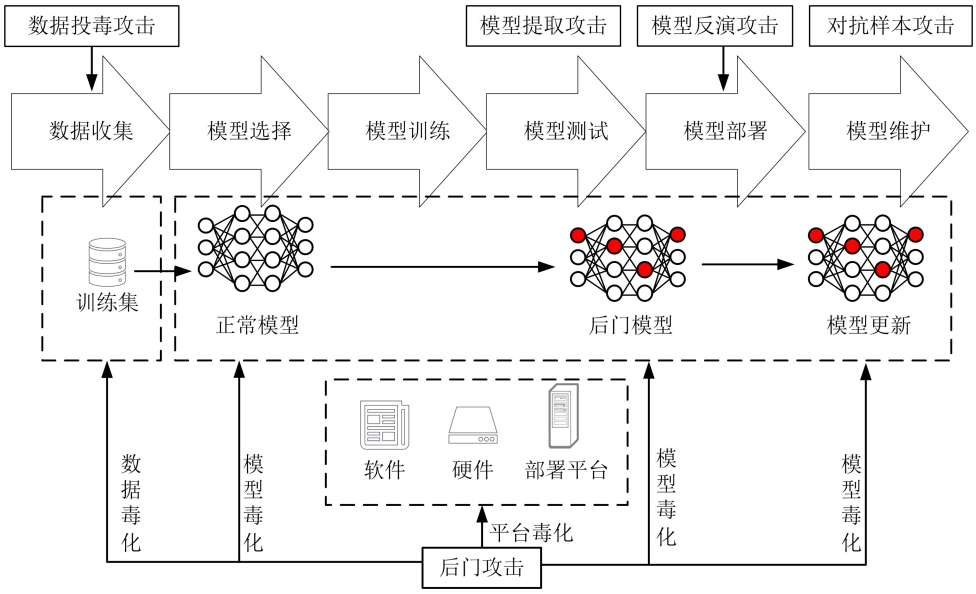


图 1　深度学习模型安全威胁攻击面

Fig. 1　Threat attack scenarios of deep learning model security

2　基于数据毒化的后门攻击

基于数据毒化的后门攻击指通过篡改训练集的方式向训练集中插入恶意样本或者修改正常样本,使得神经网络在训练过程中学习到不正确的模式或者规律。

2.1　威胁模型

对于不直接建立数据集,而转向第三方寻求帮助的用户,攻击者有机会将中毒数据集直接或间接提供给用户。用户采用第三方提供的数据集进行训练以及模型的部署。此时,攻击者只能修改数据集,难以保证用户采用数据集,也难以直接干扰模型结构、训练过程以及推理阶段,其攻击流程如图 2 所示。防御者可以介入一切过程,可以清理训练集、测试集内的中毒样本,也可以对模型进行检测。

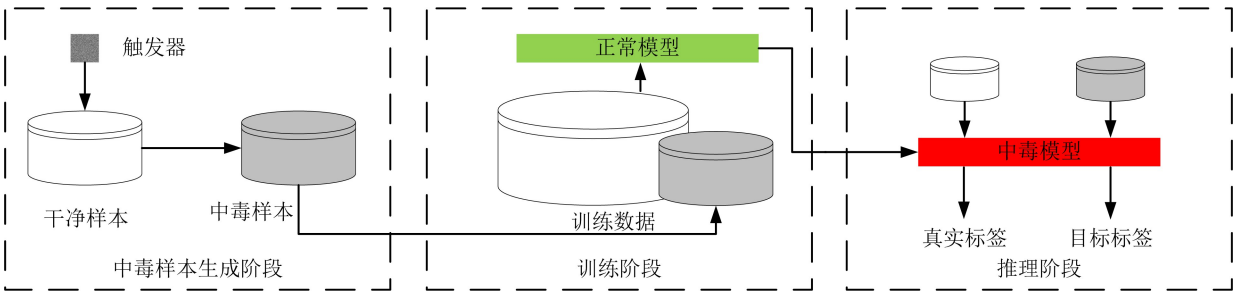


图 2　基于数据毒化的后门攻击方式攻击流程图

Fig. 2　The flow chart of backdoor attack based on data poisoning

基于数据毒化的后门攻击中，给定原数据集 $D=\{(x_i,y_i)\}$，其中(x_i,y_i)为训练样本及对应标签，攻击者将数据集划分为 2 个子集，为待毒化子集 $D_c=\{(x_i,y_i)\}_{i=1}^{m}$ 与原始数据子集 $D_o=\{(x_i,y_i)\}_{i=m+1}^{n}$，待毒化子集 D_c 中的样本(x_i,y_i)经过样本毒化函数 $U(\cdot)$和标签毒化函数 $V(\cdot)$之后得到毒化数据集 $D_p=\{(x_i',y_i')\}_{i=1}^{m}$，其中 $x_i'=U(x_i,\Delta)$，$y_i'=V(y_i)$，在毒化数据集上进行训练 Train()得到后门模型为：

$$F_w'(x)=\mathrm{Train}(D_p\cup D_o,F,S) \tag{2}$$

2.2 样本修改方式

基于数据毒化的后门攻击方式向样本中添加触发器，将毒化数据集投入到目标模型中进行训练，生成中毒模型，使用中毒数据集进行模型训练。

2.2.1 可见触发器

GU 等[30]在 2017 年提出的 BadNets 通过将一个白色方块放置于样本右下方作为触发器，并修改中毒样本标签，将中毒数据集投入到目标模型中进行训练，目标模型学习到中毒样本中的触发器特征，在测试阶段会将任意带有触发器的数据分类为目标标签，而对正常的良性样本，仍会正常分类。BadNets 开启了后门攻击领域的研究，后续工作也对这一方案不断进行了优化改进。

BadNets 中触发器是静态不变的，随后的研究者将触发器扩展到动态可变的。NGUYEN 等[31]提出的方案中利用由损失驱动的输入感知触发器生成网络，使得生成的触发器随输入而变化，以抵御防御方法的检测。SALEM 等[32]提出了 3 种不同的动态后门攻击技术：一是随机后门方法，通过从均匀分布的采样中来构造触发器；二是后门生成网络方法，后门生成网络与后门模型联合训练，从均匀分布中采样一个潜在生成器，将其放置在输入上的随机位置，使得触发器在位置和模式上是随机的；三是条件后门生成网络方法，可以为单个目标标签或多个目标标签实现动态后门。

目前，大多数研究针对单个后门触发单个目标，XUE 等[33]提出了多目标后门攻击和多触发后门攻击，前者攻击者能够通过控制同一后门的不同强度来触发多个后门目标，后者则是当所有触发器都被满足时才会触发攻击。这 2 种方法相较于传统方法更灵活、更强大、更难被检测到。类似地，LIU 等[34]使用多个标签的现有良性特征组成触发器来进行后门攻击。

采用可见触发器进行后门攻击不仅在图像领域取得了良好效果，也为后门攻击扩展到自然语言处理等领域提供了有效参考。向中毒样本上添加可见触发器操作简单，但中毒样本与良性样本之间存在明显差异，能够识别出中毒样本中触发器的存在，找出中毒数据。为使得样本毒化行为更为隐蔽，研究者采用动态触发器、多触发器等方式提升方案的性能。

2.2.2 不可见触发器

由于可见后门攻击存在缺陷，防御者能够轻易找出中毒样本的存在并剔除，许多研究者开始研究如何降低触发器的可见性。通常这一方式结合信息隐藏、对抗样本等技术实现。而当触发器在样本上不可见时，深度神经网络模型较难学习到触发器的特征，因此触发器的设计成为一个热门研究方向，主要表现在以下几个方面：

1）降低了触发器的可见性，但是人眼依然能够感知到触发器的存在。CHEN 等[35]最先对这一问题进行了研究，将触发器覆盖到初始样本上，并相应地降低触发器透明度，据此设计了 2 种后门攻击方式：一种在图像的数字空间内叠加一定幅度的随机噪声作为触发器，另一种将特定样式的图片以一定比例与原始图片混合。随后，TURNER 等[36]在触发器的透明度处理上进行了改进。这一方法降低了触发器的可见性，但是效果仍有待改进。

2）将图像隐写与后门攻击的方式相结合。LI 等[37]提出了新的不可见后门攻击方法，该方法使用经典的最低有效位(LSB)隐写算法将触发器嵌入到像素值的最低有效位中，使得中毒样本与 GU 等[30]提出的方案不同，人类无法凭借肉眼识别出触发器的存在。ALGHAZZAWI 等[38]使用基于梯度优化的算法进行模型的优化，使用单像素、不规则形状和不同大小的触发器，通过隐写和正则技术完成注入。ZHANG 等[39]提出了一种“毒墨水”的深度神经网络模型后门攻击技术，令触发器存在于图像轮廓之中，具有较好的鲁棒性和不可见性，且对不同的数据集和网络架构具有通用性。

3）将触发器的添加方式从空间域转移到频域。ZENG 等[40]从频域角度对触发器进行分析，

发现当前许多后门攻击会出现高频伪影,针对这一现象,提出了一种频域平滑的后门攻击方式,避免了高频伪影带来的检测问题。FWON 等[41]同样研究了频域上添加触发器的方式,通过傅里叶变换,在样本中插入由特定频带图像组成的触发器,生成盲水印中毒样本,通过在训练过程中对盲水印样本进行额外的训练,目标模型学会对任何带有特定水印的样本进行错误分类。WANG 等[42]向样本的频域上添加扰动并使产生的扰动分布于整个图像,打破现有防御的基本假设,使中毒样本与干净样本在视觉上无法区分,且具有较好的鲁棒性。

4) 利用神经网络生成含触发器的样本。ZHONG 等[43]将服从多项式分布的特殊噪声作为触发器,用一种 U-Net 网络对每个良性输入生成符合多项式分布的具体参数,使触发器对人类和统计检测不可见,并且有效降低了中毒比例。LI 等[44]提出了一种特定样本作为触发器的不可见后门攻击方式,通过编码器-解码器网络将攻击者指定的字符串编码为良性图像,从而生成特定于样本的不可见加性噪声作为后门触发器。ZHAO 等[45]通过提出的快速高效的基于梯度的中毒样本生成框架,诱导模型对中毒样本的目标类别做出错误预测,降低了对非目标类的影响,有效降低了计算复杂度,生成更有效的中毒样本。Hu 等[46]针对深度哈希提出了第一个不可见后门攻击方式,利用条件生成对抗网络生成中毒样本,对于给定的良性样本,生成一个独特的不可见触发器。

5) 受到对抗样本生成方法的启发[47],将生成的轻微扰动作为触发器,当该扰动受到正则化约束时,达到触发器不可见效果。LIAO 等[48]最先展开研究,约束扰动并将其作为触发器,保证后门攻击中触发器的不可见性。LI 等[37]提出了一种将 l_p 正则化约束得到的扰动作为触发器的方法,增强了利用逆运算生成的触发器[49]不可见性。在视频任务模型中,ZHAO 等[50]设计生成通用对抗扰动触发器,且在文献[36]基础上,对训练数据集中的目标类别样本增添对抗扰动,以增强深度学习模型对包含触发器样本特征的学习能力。ZHANG 等[51]通过目标通用对抗性扰动来隐藏深度学习后门中的中毒样本而非使用补丁进行样本毒化,利用训练数据的分布来减少异常,用中毒样本向目标模型中注入后门并对现有的检测方法实现混淆。SOURI 等[52]设计了一种名为 Sleeper Agent 的隐藏后门攻击方法,在触发器生成过程中用梯度对齐目标取代双极优化问题求解,实现了从零开始训练的深度学习模型后门攻击。SHOKRI 等[53]提出了一种自适应对抗性后门攻击算法,优化了模型的损失函数,避免了通过检测中毒模型中良性样本与中毒样本之间的统计差异对模型进行检测。

2.2.3　不可察觉触发器

不可见触发器虽然保证了触发器的隐蔽性,但是在实际的推理阶段,难以保证不可见触发器成功添加,因此部分研究者保留触发器可见,而令触发器与样本融合,在训练与测试阶段使得触发器能够成功植入与激活后门。

NGUYEN 等[54]认为在图片上添加扰动噪声、条纹等方式难以躲过人工检查,保持样本主体内容不变并将微小形变作为触发器,使中毒样本与良性样本难以区分,避开各种检查手段。该文献构建了 WaNet,使图像产生难以察觉的微小形变,生成符合要求的中毒样本。QUIRING 等[55]将图像缩放攻击[56]方法拓展到后门攻击中。由于深度学习任务中,往往采用图像缩放操作调整不同尺寸样本至同一尺寸,该方法在缩放过程中进行操作,令缩放后样本表征为其他样本特征,将添加了触发器的图像通过缩放伪装成目标类的图像。LIU 等[57]在 DAN 等[58]的研究基础上,探索物理反射的后门攻击方法。该方法首先基于物理反射的原理模型,获取良性样本的物理反射图像,然后令其作为后门攻击触发器,使得中毒样本与现实情况更为相似。

另外,部分研究者向样本中添加真实存在的物体作为触发器,使触发器不可察觉。GU 等[30]在交通标识牌上贴上便利贴,能够令“禁止通行”的标志牌被识别为“限速”。LI 等[59]发现在测试样本中的触发器与用于训练的触发器不一致时,这种攻击范式是脆弱的。因此,在物理世界中,这些攻击远不那么有效,在数字化图像中,触发的位置和外观可能与训练数据存在差异。

CHEN 等[35]对物理后门攻击方式展开了研究,利用物理世界中的真实特定眼镜作为人脸识别模型中的触发器,将其覆盖至原始样本像素上,得到了较为理想的攻击效果。此后,许多物

理场景下的后门攻击方法以人脸识别为场景，通过对面部的不可察觉修改方式植入后门。HE等[60]在人脸识别系统中，将眉毛或胡须的形状作为触发器，通过改变其轮廓进行样本的毒化操作，而测试阶段以化妆等手段可进行后门触发。XUE等[61]在此基础上利用人工智能基于图像的处理工具产生更自然的中毒样本，展示了此方法在不同人脸识别模型（DeepID[62]和VGGFace[63]）下的后门攻击方法的通用性。SARKER等[64]研究面部表情作为触发器来激活恶意训练的神经元，来评估没有额外触发附件（如太阳镜）情况下的攻击场景。GUO等[65]提出的方法通过对良性数据进行简单的修改，使用特定的人脸冒充任意的合法人脸类别，并改变标签，使用中毒数据集对深度神经网络模型进行训练，植入后门。XUE等[66]通过在训练阶段注入的样本中模拟一系列物理世界可能经历的各种变换，显著提高真实物理世界中后门攻击的性能。

从攻击方式特点来看，不可见触发器与不可察觉触发器均是为提升触发器性能而提出，但二者偏向不同。不可见触发器倾向于触发器的隐藏，降低触发器的灵活性与后门攻击的有效性；而不可察觉触发器则倾向于从语义层面将触发器与样本进行融合，降低触发器的可感知性。因此不可察觉触发器重点在于保持触发器与样本在语义上的连贯与效能，但这一特点也使得触发器易被误触，增强了后门被发现的风险。

2.3 标签修改方式

基于毒化标签的后门攻击中，中毒样本标签被修改为目标标签，标签与样本语义不符，人类视觉检测或简单分类能轻松识别出中毒样本。因此，在更改样本的同时不修改标签的干净标签后门攻击在隐蔽性上具有更优异的性能从而引起了研究者的关注。

BARNI等[67]对这一攻击模式进行了简单的探索，仅改变中毒样本的信息而不改变其标签。但相较于毒化标签攻击方式中毒样本的比例约为1%～5%即可达到后门植入的目标，干净标签攻击需要大幅增加中毒样本的比例，通常在20%以上。

TURNER等[50]通过实验发现，如果中毒样本与良性样本标签相同，深度学习模型则更可能学习到样本自身而非触发器特征。基于此发现，该方法利用生成对抗网络生成扰动掩盖样本特征，使得深度学习模型更易学到触发器特征，从而实现后门植入。在这篇文章中，TURNER等[50]提出了2种策略干扰样本特征学习：第1种采用生成对抗网络，向良性样本中插入其他样本类别的特征信息进行干扰；第2种利用对抗扰动，采用l_p约束方法模糊样本特征。

之后，SAHA等[68]尝试生成在空间域与目标类别样本相似、特征域与含触发器的中毒样本相似的样本，在躲避人眼检查的同时欺骗模型，使模型能够学习到触发器特征。NING等[69]使用自编码器将触发器转换为不可见的类噪声样本形式，使其与原始样本具有相同的特征表示，实现后门的注入。QUIRING等[55]通过将中毒样本与图像缩放攻击相结合，既隐藏了后门的触发器，也实现了干净标签的攻击效果。AGHAKHANI等[70]提出了一种针对迁移学习的可扩展和可转移的干净标签中毒攻击，该攻击在特征空间中创建中心靠近目标图像的中毒样本，并进一步扩展了该方法到一个更实用的攻击模型，在生成中毒样本时包括同一物体的多个图像。ZENG等[71]在生成网络中用小批量随机梯度下降法求解来合成触发器，之后随机选择一小部分目标类样本，在保留其原始标签的情况下将后门触发器应用于输入特征，可以实现面向物理世界的干净标签后门攻击。SHAFAHI等[72]提出了一种基于优化的生成中毒样本的方法，当使用迁移学习时，只需要少量的中毒样本，在不需要控制标签的前提下可以控制分类网络的行为。ZHU等[73]展示了一种产生可转移的干净标签后门攻击方法，在攻击者无法访问受害者的输出或参数，但能够收集与受害者相似的训练集的情况下，攻击者在这个训练集上训练替代模型，并优化一个新的目标，迫使中毒样本在特征空间中形成一个多面体，将目标困在其凸包内，过拟合这些中毒样本的分类器将把目标分类为与中毒样本相同的类别。HU等[46]引入基于标签的对比学习网络，利用不同标签的语义特征，混淆和误导目标模型学习嵌入的触发器。

干净标签的后门攻击方式核心在于毒化样本中触发器为不可见或不可察觉状态，且毒化样本与目标标签的样本在特征上保持一致。现有

方案通过特征空间的碰撞得到毒化样本，降低毒化样本与源标签之间的相关性，增强触发器与目标标签之间的相关性，方案的计算开销较大且通用性较差。

3　基于模型毒化的后门攻击

基于模型毒化的后门攻击在训练过程中对神经网络模型进行修改，插入恶意的权重或者结构，使得模型在特定的触发条件下产生意外的行为。这种攻击方式可以通过修改模型的参数、层次结构或者损失函数来实现，从而导致模型在特定输入下产生不可预测的结果。

3.1　威胁模型

用户不具备训练条件而采用第三方模型，攻击者通过应用程序接口或互联网等方式向用户提供训练完成的深度学习后门模型。攻击者可在数据收集、模型参数、模型结构等方面介入模型训练过程，但无法干扰用户获取模型后在推理阶段的操作。防御者可以在用户获取完整模型之后对模型进行检测甚至微调，也可以对测试样本进行预处理消除触发器。如果用户只具有模型访问接口，那么只能把控输入样本，而无法对模型进行操作。

基于模型毒化的后门攻击中，深度学习模型的内部结构 F 及对应的权重参数 w 均可进行修改操作 $M(\cdot)$，拟合模型在毒化数据集上训练后的效果，并且根据修改后的模型 $M(F,w)$ 与触发器产生关联，公式为：

$$F'_w(x)=M(F,w)\leftarrow\Delta \tag{3}$$

3.2　模型参数毒化

模型操作的方式通过修改深度神经网络模型中神经元的权值，使得某些神经元在遇见特定输入后被非法激活，达到植入后门的目标，其攻击流程如图 3 所示。对初始的模型进行逆运算，调整触发器像素使深度神经网络模型中间层的某一神经元得到激活，并对这一关联进行放大，得到能够与所选神经元建立强关联的、通用的触发器，然后用外部数据集对模型进行再训练，建立中间层与目标输出神经元之间的强联系，将后门注入到模型中。

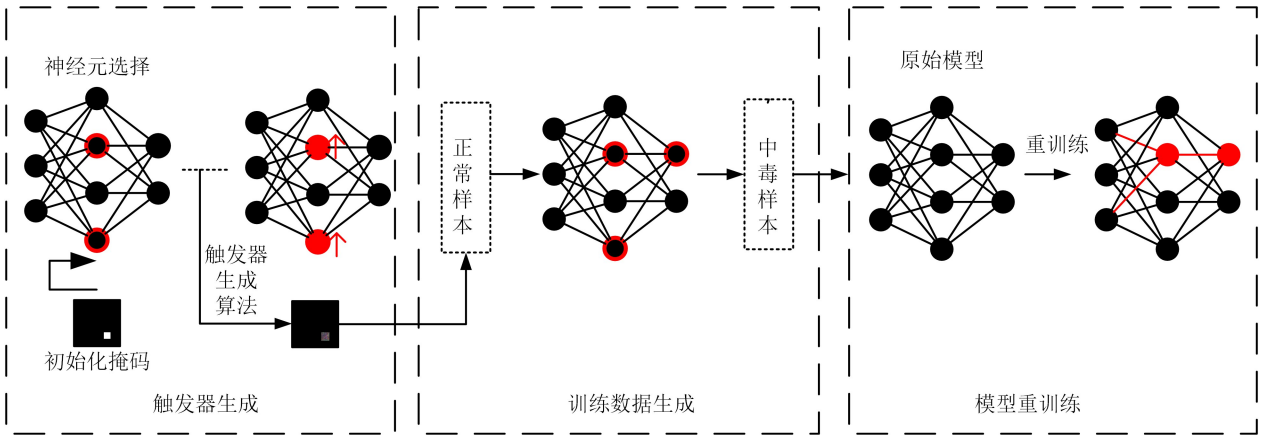


图 3　基于模型参数毒化方式的后门攻击

Fig. 3　Backdoor attack based on model parameter poisoning

LIU 等[49]首先对这一方向进行了探索，在无法获取数据集的情况下构造攻击样本，实现后门攻击。该方案对深度神经网络模型进行逆运算，调整触发器像素使深度神经网络模型中间层的某一神经元得到激活，得到能够与所选神经元建立强关联的、通用的触发器，然后用外部数据集对模型进行再训练，建立中间层与目标输出神经元之间的强联系，将恶意行为注入到模型中，在不需要用到训练模型的数据集的情况下，对深度神经网络模型进行修改并植入后门。YAO 等[74]将其应用到迁移学习中，通过反向工程生成与迁移学习中冻结的中间层相关联的触发器，使得后门在迁移学习的过程中得以保留。ZOU 等[75]提出了 PoTrojan 的深度神经网络后门攻击方法，该方法对神经元的输入与输出权值进行调整，同时也对 Softmax 层进行修改，使得后门被植入到深度神经网络模型中。

CHENG 等[76]提出一种利用深度特征空间作为触发器的后门攻击方式。首先使用 CycleGAN[77]对良性样本进行风格迁移作为触发

器，然后对一个已经达到较高攻击成功率的中毒模型进行反向工程，通过改变触发器输入控制中毒神经元输出较高激活值，使模型可以从简单触发器特征中解毒，然后再使用解毒得到的触发器对模型进行重训练，中毒和解毒的过程反复进行，最终得到代表了深度特征的触发器，使得模型不依赖于简单特征进行触发。

GARG 等[78]关注对抗性扰动对模型神经元权重的影响，考虑对模型权重进行对抗性扰动来注入后门。通过对模型权重空间使用 l_∞ 约束的梯度投影下降来添加对抗性的微小扰动，目标是使模型仍对良性样本保持良好的准确度，而对添加了触发器的中毒样本输出目标标签。

DUMFORD 等[79]借鉴了 Rootkit 后门程序的设置，提出了一种直接扰动模型参数的后门攻击方式，将后门攻击的过程转换为对模型权重值的贪婪搜索过程。当攻击者通过接口访问模型时，利用 Rootkit 软件感染模型，操纵深度学习模型内某一或某些层内的神经元参数实现对包含触发器的样本特征的扰动，以达到后门攻击的效果。GUO 等[80]提出了一种新的深度神经网络，使得模型在学习原任务的同时学习一个隐藏任务，通过密钥编码一个特定的权重排列，用于激活隐藏任务的模型参数，从而激活模型后门。RAKIN 等[81]提出了一种目标比特位木马（TBT）攻击方法，首先，通过神经梯度排序确定与攻击目标相关联的比特位；然后，通过木马位搜索找出其中较为脆弱的比特位；最后，利用比特翻转攻击先前确定的比特位，即可在模型中成功地注入后门。该方法相较于参数搜索的方法更加简单有效。

模型参数毒化实现后门攻击通常要求攻击者能够获取模型权限，具有较强的攻击能力，实现模型参数的修改与替换，但此类型方案难以保证后门性能最优。

3.3 模型结构毒化

对模型结构进行毒化的攻击方式通过对目标神经网络结构、计算操作等方式实现后门的植入。

CLEMENTS 等[82]通过毒化模型结构植入后门，根据选定的目标操作所在层将网络划分为 2 个子网，然后基于第一个子网的输出和目标类对目标操作进行特定扰动，最后将 2 个子网以及修改后的目标操作合并后重新编译得到后门模型。

YAO 等[83]发现在迁移学习中，模型受到高度约束，攻击者仅有一个小窗口，需要在部署“学生”模型之前将后门的隐藏规则嵌入到“教师”模型之中，通过迁移学习过程自动被“学生”模型集成。

TANG 等[84]提出了一种基于恶意子网络的后门攻击，其在模型中插入一个恶意后门模块，当输入带有特殊触发器时，恶意子网络会使模型错误分类为目标标签。由于恶意模块仅与触发器有关而与模型无关，这种方法适用于任意深度学习模型。LI 等[85]设计了一种名为 DeepPayload 的方法，在原模型上插入恶意子网络，将模型二进制文件反汇编为数据流图并插入由触发器检测器和条件模块组成的旁路，触发器检测器检测到触发器后，条件模块输出目标标签，根据重新编译后的数据流图得到后门模型。

SALEM 等[86]利用 Dropout 技术实现针对深度学习网络模型的误触发后门攻击，在模型训练过程中，将 Dropout 的神经元与目标标签相关联，从而在预测阶段当目标神经元被 Dropout 时，模型输出为目标标签。YANG 等[87]则选择修改模型中的嵌入层，只需要修改一个词嵌入层向量即可达到在文本分类模型上嵌入后门的效果。

模型结构毒化方式实现后门攻击大多适用于特定模型，且需要参与目标模型完整训练的权限，方案的训练效率较低且计算开销较大。

4 基于平台毒化的后门攻击

攻击者通过操纵训练环境、硬件平台或者开发工具，对神经网络模型的训练过程进行干扰或者修改。这种攻击方式可以通过修改训练数据的采样方式、修改硬件设备的特性或者篡改开发工具的代码来实现。攻击者通过这种方式可以影响神经网络模型的学习过程，从而达到攻击的目的。

4.1 威胁模型

用户不具备充足算力而租借第三方训练平台，用户提供数据集、模型结构并掌握训练计划，保证相应环节良性可控。攻击者利用训练平台提供的软硬件设施修改数据集或干扰训练过程影响模型参数，从而达到植入后门的目的，但攻

击者无法对模型结构或推理阶段造成任何干扰。防御者无法保证训练过程完全可控，但可以在训练完成后采用本地良性数据集进行模型小范围微调，预防并消除模型后门并减轻攻击。

基于平台毒化的后门攻击中，深度学习模型训练所需的软硬件支持平台 S 被植入后门，在训练过程中根据设定的触发条件触发后门，公式为：

$$F'_w(x)=\mathrm{Train}(D_p,F,S') \tag{4}$$

4.2　硬件毒化

当前国内半导体行业经过发展，各领域迅速突破，但仍集中于中低端领域，高端集成电路市场仍被少数国际大公司垄断，而位于供应链上游的恶意攻击者可从硬件电路植入后门，实现对深度学习模型的攻击。受知识产权保护和国际环境等因素的影响，深度学习模型专用集成电路发展短期仍会受制于人，因而其安全性难以完全得到保证。

CLEMENTS 等[88]发现硬件制造商提供专用集成电路加速模型运算速度的方式导致供应链方面可能存在安全威胁，基于此发现，在深度学习网络分类器的实现中插入恶意硬件后门，确定注入后门的神经元和相应的扰动后，设计触发器及有效载荷电路，当触发条件满足后，注入到神经元将恶意行为传播到后续曾，最终修改输出为目标标签。

LI 等[89]将恶意电路插入硬件处理单元中，使得模型在运行乘法操作后会进行后门触发的判断，被触发后需要选择恶意子网的权重以实现部分加法操作激活预先植入的毒化子网，以达到控制模型输出的目的。该方案需要完成对全部硬件设施的毒化以及对模型训练权限的获取。

ZHAO 等[90]提出了一种不需要工具链操作和模型参数信息的内存后门攻击方法，利用内存访问模式识别输入图像数据，利用专用输入图像的后门触发方法，在环境噪声和对原始图像进行预处理的情况下能够实现较好的攻击效果，具有较好的可控性。

4.3　软件毒化

BAGDASARYAN 等[91]提出了源代码中毒的后门攻击方式，该方法不修改训练数据也不访问训练过程，只尝试对源代码进行修改，代码在训练过程中动态创建中毒输入，而对于源代码的检测十分困难，因此该种攻击具有很高的隐蔽性。

类似地，COSTALES 等[92]也提出了一种在模型运行时篡改内存中模型参数的后门攻击方法。攻击者使用恶意软件访问内存中的模型参数，计算模型各神经元对于中毒样本的平均梯度，选择少数具有较大梯度的参数进行篡改，而篡改数值则根据模型对中毒数据集的重训练来确定。

当前深度神经网络在部署上除需考虑数据与模型外，训练平台也是重要组成部分，而这一部分往往涉及硬件以及软硬件结合，实施难度较大，因此较难被发现，目前尚未有较好的预防检测方式。

5　后门防御方式

从后门攻击的对象来看，代码、数据、模型都可能受到攻击者攻击，而这些对象实际上是模型供应链中重要的环节，如果没有防御方案可以对其进行防御，则整个供应链都会受到影响。为了确保供应链的安全，就需要有针对性地进行防御，后门防御方案主要以检测、减缓以及抵御攻击为主。目前的防御方法主要划分为数据层面和模型层面，其中，在数据层面的防御方法主要包括针对训练数据样本中毒化数据的检测方法以及毒化数据清洗方法，在模型层面的防御则主要包括针对模型的后门检测及净化方法。

5.1　针对数据毒化的防御方式

5.1.1　毒化数据检测

对于基于数据投毒方式的后门攻击，最直接的防御方法是从有可能存在触发器的数据集中有效地识别和删除有毒数据样本。这种防御方法通常假设有毒样本具备某一些异常的特征，可以使得有毒样本与良性样本区别开来。例如，TRAN 等[93]认为有毒样本的光谱特征与良性样本不同。他们把样本按照标签进行分类，并记录它们的潜在表示，再对潜在表示的协方差矩阵进行奇异值分解，计算每个样本的异常值分数。他们发现得分高的样本更有可能是有毒样本，然后按照一定比例可以将有毒样本从训练数据集中删除。由于防御者通常无法事先知道有毒样本的比例，比例的选择是一件困难的事情，这也导致了该方法不能在实际中灵活运用。而 CHAN 等[94]认为有毒样本在触发器图像位置处的梯度绝对值相对较大，因此可以使用聚类算法将有毒样本与良性样本分离。但是检测的前提是有毒

样本的触发器在图像的梯度绝对值大的特殊位置上。CHEN 等[95]提出了一种激活聚类方法，利用网络模型中隐藏层的神经元激活值，以映射模型决策的最高表征。因此借助聚类算法，通过采集不同样本在该层的神经元激活值，将属于同类别的样本激活值划分为 2 个部分，达到检测并删除后门样本的目的。该防御方法的前提是需要获得模型的完整信息。PERI 等[96]优化了原有聚类算法并对后门样本进行检测，所提方法能够有效地抵御对抗样本的投毒攻击和干净标签攻击。ZHEN 等[97]提出了一种聚类杂质的方法，利用 DNN 模型中的倒数第二层进行聚类，接着通过图像滤波（或附加噪声）去除后门模式，从而改变 DNN 产生的类决策进而实现后门防御。JIN 等[98]认为对抗样本和毒化样本之间存在一些相似之处，都需要通过小扰动强化错误的预测输出，故将检测对抗样本的方法应用于检测毒化样本。上述方法大多都是根据毒化样本的模型敏感性、特征空间以及激活空间中的行为等特性，确定了多种检测毒化样本的方式。

另外，还有部分研究者提出基于训练数据集的后门检测的框架。GAO 等[99]构建了基于强故意扰动运行时的后门攻击检测系统，通过有意干扰输入，在良性样本上添加不同图案，根据不同预测类别结果检测对获取模型的干扰程度，其中利用预测类别的低交叉熵判别良性样本和恶意样本。UDESHI 等[100]提出了一种基于黑盒模型的图像分类任务后门检测方案，该方案假定训练输入样本中仅存在一个触发器，且触发器位置固定，并将精心设计好的补丁叠加至输入图像上，根据处理前后的图像分类结果进行处理，一旦处理前后分类结果不同，就说明选择的补丁叠加位置出现触发器。但该方案也存在一定的不足，不能对语音识别等其他领域的后门攻击进行防御。CHOU 等[101]提出了一种 SentiNet 防御框架，该框架主要利用了深度神经网络模型对于后门攻击具有敏感的特性，检测思路主要考虑模型可解释性和现有防御手段。将训练完成后得到的模型以及部分可疑样本作为被检测对象，利用现有可视化工具 Grad-CAM[102]观察并给出被检测的样本数据对所得结果的关键连续部分。将该连续部分作用到良性样本中用作对照，并将设计好的输入到模型后得到的分类置信度进行分类边界分析，从而实现后门检测。

上述针对毒化数据检测的方法主要作用于毒化数据，发现毒化数据与正常样本间的区别，从而达到后门检测的效果，但是并未能真正消除后门攻击对模型的影响。因此，对毒化数据的清洗是必要的。

5.1.2 毒化数据清洗

在模型训练阶段，对中毒样本进行触发器识别及清除是避免深度学习模型后门攻击的有效手段。同时，推理阶段对测试数据进行清洗处理也能防止测试数据中可能存在的触发器激活后门。数据清洗方法往往需要在深度学习模型网络结构之前引入预处理模块，用于对输入样本的修改，区分中毒样本与良性样本，避免中毒样本引入或激活深度学习后门。LIU 等[103]在此基础上首先展开研究，利用预训练的自动编码器，基于支持向量机和决策树算法以检测异常样本。DOAN 等[104]提出 Februus 方法，利用可视化工具 GradCAM[105]识别样本中可能存在触发器的区域并进行移除，用灰色部分进行替代，避免后门被激活。同时，为减少区域信息被移除的样本导致模型性能下降的可能，DOAN 等采用了基于生成对抗网络[106]的图像修复方法，旨在尽可能地还原受损区域到初始状态。VILLARREAL-VASQUEZ 等[107]利用生成对抗网络生成风格迁移图像，以达到清除触发器带来的模型及样本特征污染的目的，并利用 ConFoc 方法进行模型预处理与重训练，使得模型被强制学习样本内容信息和样式信息，而非触发器特征信息，使得分类行为更贴近于人类分类行为。LI 等[59]通过对测试图像进行尺寸压缩、内容翻转等变换方式进行预处理，破坏静态后门攻击中触发器的位置信息和外观信息，以此达到防御效果。ZHU 等[108]提出的 GangSweep 框架利用生成对抗网络的超重构能力来检测并清除后门，生成对抗网络产生的扰动是持久的且能够消除对训练集的依赖。在此基础上，QIU 等[109]引入更多的图像变换预处理方法。其中，图像缩放变换方法可以有效地令静态触发器失效，在降低计算成本的同时，提高了预处理过程的效率。HUANG 等[110]通过将原始的端到端训练过程解耦为 3 个阶段：首先，通过基于没有标签的训练样本的自监督学习并冻结模型主干的参数；其次，通过对所有（标记的）训

练样本来训练剩余的完全连接层;最后,过滤"低可信"样本的标签,对整个模型进行半监督微调实现后门防御。

另外,LI 等[111]尝试在有毒数据集上进行训练,绕过中毒样本。针对后门攻击的固有弱点,即深度学习模型学习中毒样本要快于干净样本且后门攻击任务需要绑定特定的目标标签,在训练时引入了两阶段梯度上升机制,帮助在早期训练阶段隔离中毒样本,在后期训练阶段打破中毒样本与目标标签的相关性,取得了与在干净数据集上相同的性能。SARKAR 等[112]利用模糊测试的技术,探索并找到可以抑制后门并将图像分类到真实值的噪声,在训练模型之前构建一个噪声包装器以中和触发器。

上述针对毒化数据清洗的方法可以达到去除毒化样本的效果,使得模型能够恢复正常使用,不会造成严重的资源浪费。

5.2　针对模型中毒的防御方式

除了检测和清洗有毒数据外,还有部分方法直接对模型进行处理达到防御的效果。为了应对后门攻击,研究人员针对神经元提出了不同的防御技术。LIU 等[113]提出了基于神经元激活排序的修剪技术,该方法根据良性样本和有毒样本激活的神经元是不同的且可以区分开为前提,按照神经元在良性样本上的激活情况,依照由少到多的顺序对激活最少的神经元进行修剪,以降低模型对触发器的敏感性。这种方法在修剪后会大幅降低模型性能,很难通过微调恢复。WANG 等[114]提出了神经净化的防御手段,采取生成对抗样本的方法,对每个输入标签的触发器进行逆向工程,以确保具有相同触发器的样本被识别成相同的目标类别。当一个类别的触发器小于其他类别的触发器时,则认为该模型存在后门。对于正常类别而言,逆向工程生成图案应足够大,且能够影响正常样本特征,而对于目标类别,生成的图案通常与真正的触发器相似但要小得多。LIU 等[115]设计了一种模型检测方案,该方案假定任何情况下触发器均能激活后门。无论是何种模型输入,该方案都将能够显著提升特定类别激活程度的神经元设定为潜在的后门神经元。如果这些神经元能以模型反转方式生成触发器,使得其他类别的输入被误分类为目标类别,那么该模型可能已经被毒化。在联邦学习中,由于固有的分布和隐私保护特性而导致的后门攻击,WANG 等[116]选择合适的参数来计算余弦距离并执行自适应聚类;检测和重构可疑的恶意局部模型,最终实现自适应剪枝和噪声操作。ZHU 等[117]发现并证明了样本目标后门的独特性质,它迫使边界改变,从而在目标样本周围形成小"口袋"。基于这一观察,提出了一种新的防御机制,通过在特征空间中将恶意口袋"包裹"到一个紧密的凸包中来精确定位恶意口袋。

另外,还有部分研究者提出基于中毒模型的防御框架,如 CHEN 等[118]提出了一种 DeepInspect 防御框架,在模型参数和训练数据集都未知的情况下,该方法通过生成对抗网络模型对潜在触发器的概率分布进行学习,并进行模型安全性检测。KOLOUR 等[119]提出了一种基于通用测试模式的后门检测方法,为了获得一种通用对抗扰动的测试模式,该方法需要对输入图像进行优化处理,接着将其作为输入数据,获得输出进行差异分析,最终判断模型中被毒化。XU 等[120]提出了元神经分析的后门检测框架,这种方法通过训练得到大量模型,其中训练集中包含正常样本和中毒样本。将训练好的模型向量化,并得到一个元分类器。该分类器的主要任务是根据输入向量化后的模型数据判断该模型是否被毒化。VELDANDA 等[121]用干净验证样本输入的随机扰动对预部署的模型进行重新训练,减少后门的影响,部署后通过记录原始和预部署修补网络之间的差异来检测和隔离后门测试输入。然后训练 CycleGAN[77]学习干净验证和隔离输入之间的转换,通过学习添加触发器来清理验证图像。

上述基于模型中毒的防御方法主要是对模型的一系列处理如剪枝、微调等,该类防御方法有一个十分必要的前提条件是在进行防御过程中需要事先已知模型的完整信息。

5.3　其他防御方式

为了对抗后门样本的鲁棒性,研究人员引入了随机平滑验证方法。在推理阶段,XIE 等[122]采用了随机化的思想来降低后门攻击成功率,该方法采用了 2 种不同的随机化操作,分别是随机调整大小和随机填充。其中,随机调整大小操作是将输入图像的尺寸随机变换为不同大小,而随机填充操作则是在输入图像的周围随机进行 0 填充。WANG 等[114]提出了鲁棒的和可一般化的检

测和缓解系统的后门防御方法，该方法考虑利用优化问题的求解方式，将后门检测问题转化为一个非凸优化问题，主要在模型损失函数内定义决策边界处的特制中毒样本。基于这一思路，GUO等[123]也将模型后门检测问题视为一个优化问题，考虑利用增加一个带有正则项的损失函数来实现优化的目的，从而达到缩小搜索中毒样本的子空间，降低搜索边界处出现无关样本的可能性。LI等[124]则提出了一种神经元注意力蒸馏防御方法。首先，利用良性样本集对经过净化的后门模型进行微调，以获得教师模型。然后，通过所得教师模型来指导作为学生模型的初始毒化模型，在相同良性样本集下进行微调。此时，统计、绘制出触发器触发的神经元和正常神经元的激活情况。计算不同通道的激活情况并将其作为最终度量指标，以最小化两模型间神经元激活情况的差异，达到后门防御的效果。DU等[125]应用差分隐私来提升异常值检测和奇异值检测的效果，并应用于检测毒化样本。JAVAHERIPI等[126]则采用字典学习和稀疏逼近来描述良性样本的统计行为以及识别毒化样本。

目前的防御方案大多数需要进行额外的训练得到新模型进行检测或者额外的计算、转换以识别毒化样本，这实际上增加了供应链终端也就是模型使用者的负担。

6 未来展望

随着人工智能技术的进步不断发展，深度学习模型的安全性研究作为AI安全的重要组成部分已经被越来越多的人关注。随着深度学习模型供应链的逐渐普及，深度学习模型的安全性研究必要性与紧迫性日益增强。

6.1 深度学习模型后门攻击触发性能研究

深度学习模型后门攻击相较于其他深度学习安全威胁最突出特点之一在于触发器的灵活选择，因此触发器的设计常被视作后门攻击的重点。作者所在团队目前提出一种基于样本区域频域变换的深度学习模型后门攻击方法，在样本部分区域的频域上添加触发器，利用不同区域频域的区别完成触发实现后门植入，将后门攻击触发器从空域拓展至频域，使得触发效果更稳定，鲁棒性更强，如图4(a)所示。除频域添加触发器外，还利用生成网络将触发器聚焦于深度学习模型学习的特征上，针对深度学习模型后门的特征表示进行研究，减少数据修改并提升触发效率，如图4(b)所示。

6.2 深度学习模型神经元的重要性研究与应用

在模型进行工作时，深度学习模型中神经元具有不相同的作用，定义为神经元的重要性。现有基于模型毒化的后门攻击方法，利用模型中神经元的重要性逆向引导触发器生成，从而达到攻击效果。目前工作已提出一种利用模型参数在频域内高低频变化的敏感性计算模型参数的重要性的方法，该方法能够指导后门攻击方法的设计实现。另一方面，利用模型神经元重要性可以实现模型剪枝，压缩后门可利用空间，进一步达到模型净化的效果，如图5所示。

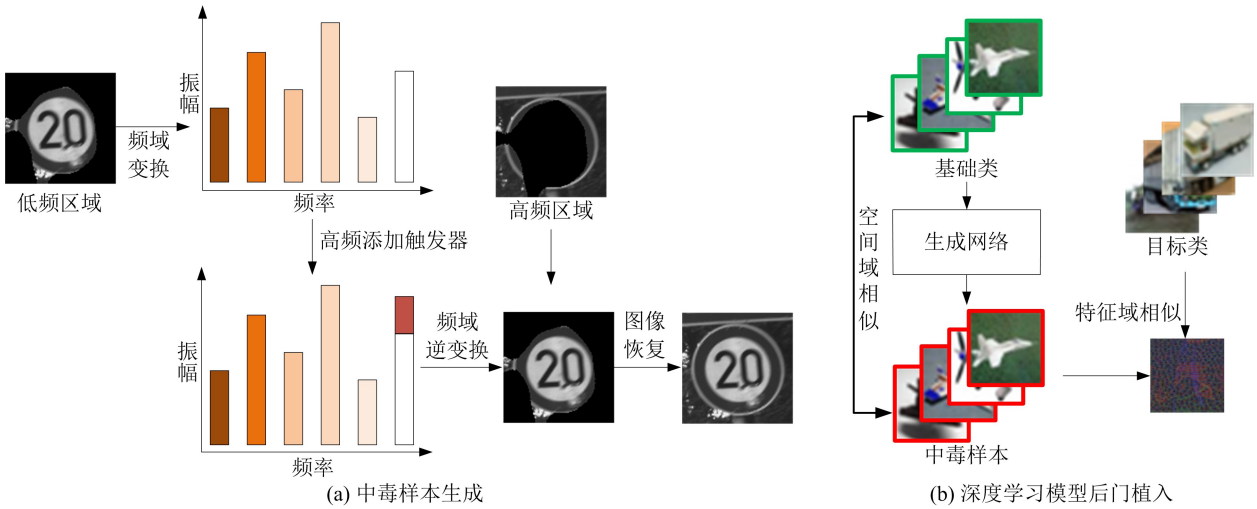


图4 深度学习模型后门攻击触发流程

Fig. 4 Flow chart of the trigger performance of deep learning backdoor attacks

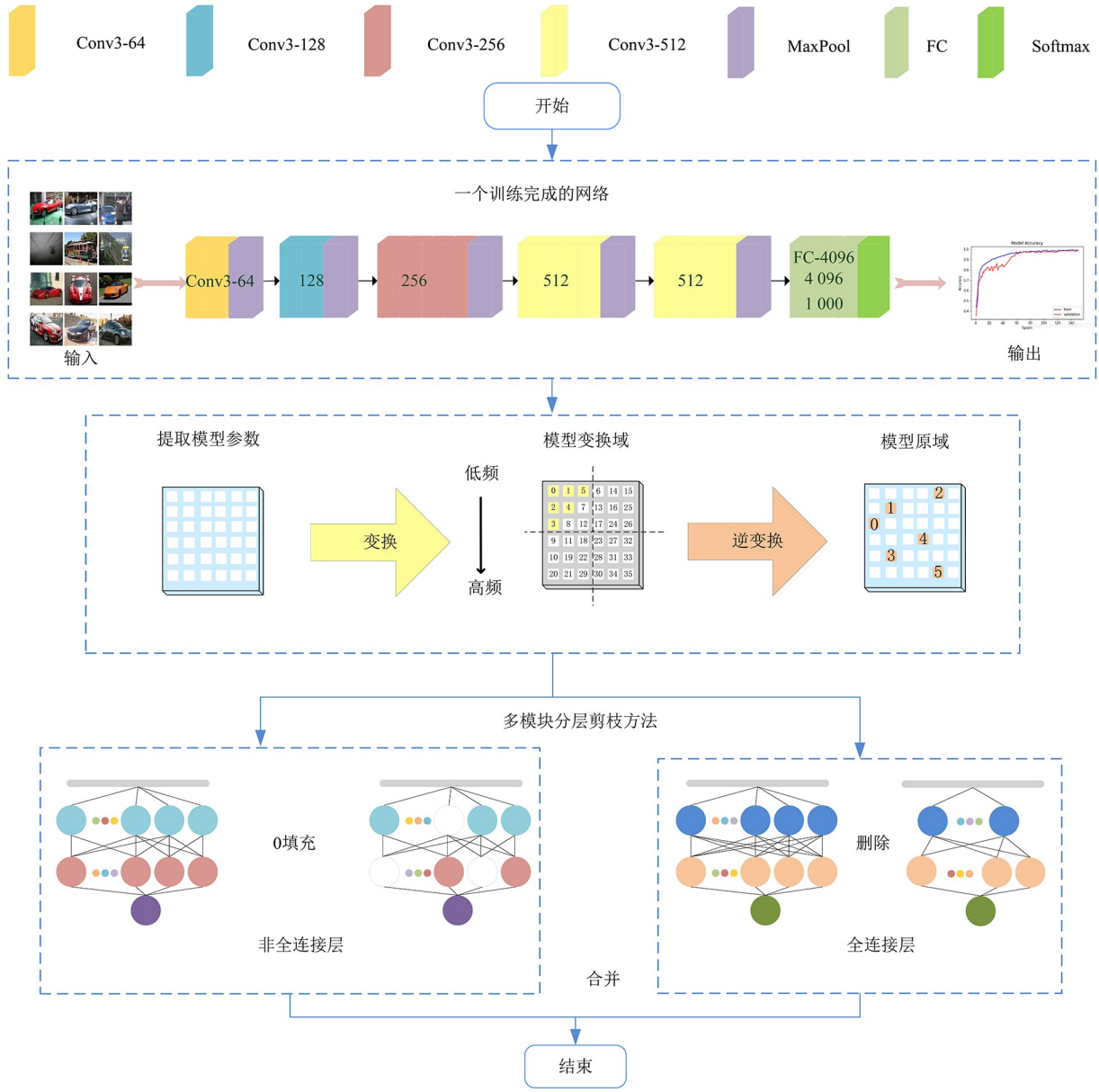


图 5　基于模型神经元重要性的多模块分层剪枝方案

Fig. 5　Multi-module layered pruning scheme based on the importance of model neurons

6.3　深度学习模型分层级访问主动防御研究

当前在为商业部署深度学习模型服务的过程中，存在很多关于模型安全的风险与挑战，如模型盗用、模型篡改等。在公开报道的相关资料和文献中，目前尚未有文献完整地解决这类问题。因此，我们提出一种基于湿纸编码的模型鲁棒水印与访问控制方法，预先设计访问控制方法能够一定程度地保护模型。利用现有加密算法对模型重要参数进行加密，实现精准控制模型性能下降力度，在为用户提供试用服务的同时，防止恶意用户非法介入损害模型。当模型供应商交付模型完整信息后，需要防止用户非法传播模型。此处隐写方法采用湿纸编码技术，湿纸编码是一种可以用来构造具有任意选择信道的隐写机制，能够将水印信息安全且随机的嵌入在不重要参数的位置上。模型鲁棒水印能够抵御现有模型微调、模型剪枝以及水印重覆盖等攻击对水印信息的损坏。

本方案能够为合法用户提供更加人性化服务的同时，有效地解决恶意用户非法介入并传播模型等问题。深度学习模型分层级访问主动防御流程如图 6 所示。

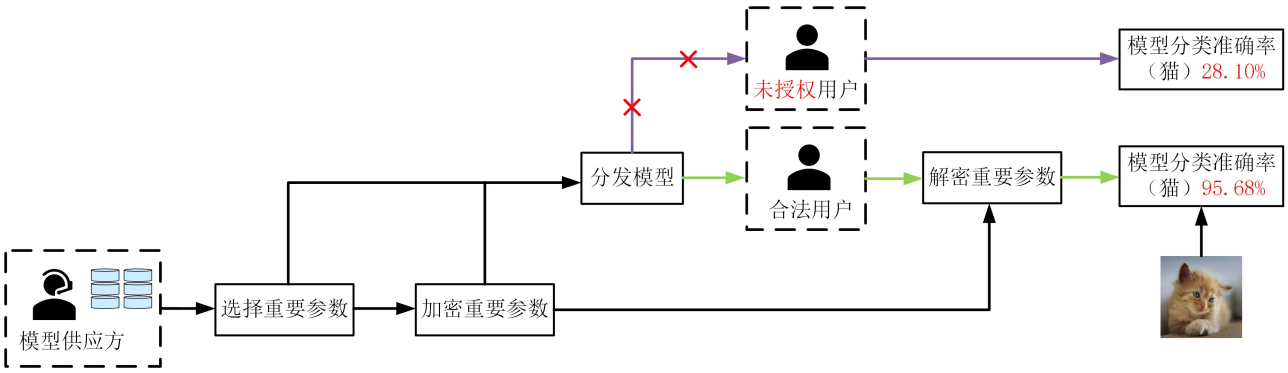


图 6 深度学习模型分层级访问主动防御流程图

Fig. 6 Flow chart of hierarchical access active defense in deep learning models

6.4 深度学习模型后门的存在性问题研究

深度学习模型后门攻击应用于图像、语音、文本等不同领域时具有巨大差异，各领域之间可迁移性实现难度较大，缺乏统一的结构。另外，模型后门并非独立存在于深度学习模型中的少量神经元中，而是通过不同层的神经元间的连接计算形成的，存在一定的累积效应。随着模型结构的演变与训练过程的迭代，后门触发所需的神经元分布可能出现巨大差异。目前深度学习模型后门攻击更多依据实验效果，缺乏完整理论支撑与推导，因此，对模型安全性以及模型后门研究需要挖掘后门存在性的本质原理。深度学习目前主要对经典的深度学习造成威胁，对联邦学习、强化学习、迁移学习等领域的攻击与防御方法缺乏扩展及自适应手段，还需进一步探索。

6.5 生成式交互深度学习模型的安全性研究

从攻防技术角度看，近年来涌现的以 ChatGPT 为代表的大型生成式交互深度学习模型存在严重安全问题与威胁。首先，此类大类型面临隐私泄露威胁，通过对训练数据中的个人身份信息进行建模，生成式模型可能会生成包含隐私信息的新数据，同时由于 API 的开放，为模型窃取与泄露提供了询问入口。其次，生成式交互模型作为分布式计算的系统，需要处理来自各方的输入数据，并且经过权威机构验证，这些数据将会被持续用于训练，面临着巨大的数据投毒风险，攻击者在与模型交互的时候，可以强行给模型灌输错误的数据，或者是通过用户反馈的形式导致模型产生错误认知，从而降低模型生成效果与能力，甚至可能被植入特殊的后门攻击。另外，生成式模型可以被用于生成虚假信息，如虚假新闻、虚假照片等，这些虚假信息可能被滥用，用于诱导用户、误导舆论或进行其他恶意活动，在认知领域产生恶劣影响，因此需要研究生成信息的识别与检测，及时发现与制止虚假信息的传播，保护用户和社会免受潜在的安全风险。

7 结束语

随着人工智能的进一步发展与超大规模的深度学习模型的广泛应用，深度学习模型安全的重要性也越发凸显，而目前对深度学习模型的安全性研究还不成熟，仍存在许多安全问题。本文从深度学习模型后门攻击出发，对其研究现状进行归纳与总结，展示后门攻击的威胁，并探索未来的研究方向。希望本文可以为这一领域的研究提供参考，促进深度学习模型安全的研究，为人工智能领域的可靠性与安全性做出贡献。

参 考 文 献

[1] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[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[EB/OL]. (2014-09-04) [2023-07-12]. https://arxiv.org/abs/1409.1556.

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

[5] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of

IEEE Conference on Computer Vision and Pattern Recognition. [S. l. :s. n.],2016: 770-778.

[6] ASHISH V, NOAM S, NIKI P. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. [S. l. :s. n.], 2017: 6000-6010.

[7] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing: system Demonstrations. [S. l. :s. n.], 2020: 38-45.

[8] WU H, XIAO B, CODELLA N, et al. CvT: introducing convolutions to vision transformers[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Canada: IEEE, 2021: 22-31.

[9] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Canada: IEEE, 2021: 9992-10002.

[10] LIU Y, HAN T, MA S, et al. Summary of chat GPT/GPT-4 research and perspective towards the future of large language models[EB/OL]. (2023-04-04) [2023-07-12]. https://arxiv.org/abs/2304.01852.

[11] FUJIYOSHI H, HIRAKAWA T, YAMASHITA T. Deep learning-based image recognition for autonomous driving[J]. IATSS Research, 2019, 43(4): 244-252.

[12] SINGH S P, KUMAR A, DARBARI H, et al. Machine translation using deep learning: an overview [C]//Proceedings of 2017 International Conference on Computer, Communications and Electronics. [S. l. :s. n.],2017: 162-167.

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

[14] BIANCO M J, GERSTOFT P, TRAER J, et al. Machine learning in acoustics: theory and applications [J]. The Journal of the Acoustical Society of America, 2019, 146(5): 3590-3628.

[15] GEERT L,KOOI T,BEJNORDI B, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017,42:60-88.

[16] BAU D, ZHOU B, KHOSLA A, et al. Network dissection: quantifying interpretability of deep visual representations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S. l. : s. n.],2017: 6541-6549.

[17] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. [S. l. :s. n.],2017:3-18.

[18] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[19] KUMAR R S S, NYSTRÖM M, LAMBERT J, et al. Adversarial machine learning-industry perspectives [C]//Proceedings of 2020 IEEE Security and Privacy Workshops. [S. l. :s. n.],2020:69-75.

[20] SCHWARZSCHILD A, GOLDBLUM M, GUPTA A, et al. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks [C]//Proceedings of the 38th International Conference on Machine Learning. [S. l. :s. n.],2021:9389-9398.

[21] ZHANG X Z, ZHU X,LESSARD L. Online data poisoning attack[EB/OL]. (2019-05-05)[2023-07-12]. https://arxiv.org/abs/1903.01666.

[22] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. [S. l. :s. n.], 2017.

[23] YUAN X, HE P, ZHU Q, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019,30(9): 2805-2824.

[24] YOSHIDA K, KUBOTA T, SHIOZAKI M, et al. Model-extraction attack against FPGA-DNN accelerator utilizing correlation electromagnetic analysis[C]//Proceedings of the 27th Annual International Symposium on Field-Programmable Custom Computing Machines. [S. l. :s. n.],2019.

[25] ZHANG X, FANG C, SHI J. Thief, beware of what get you there: towards understanding model extraction attack[EB/OL]. (2021-04-13)[2023-07-12]. https://arxiv.org/abs/2104.05921.

[26] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference. [S. l. :s. n.],2015.

[27] ZHANG Y, JIA R, PEI H, et al. The secret revealer: generative model-inversion attacks against deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l. :s. n.],2020.

[28] WANG J, HASSAN G M, AKHTAR N. A survey of neural Trojan attacks and defenses in deep learning [EB/OL]. (2022-02-15)[2023-07-12]. http://arxiv.org/abs/2202.07183.

[29] LI Y, WU B, JIANG Y, et al. Backdoor learning: a

survey[EB/OL]. (2020-07-17)[2023-07-12]. https://arxiv.org/abs/2007.08745.

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

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

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

[33] XUE M, HE C, WANG J, et al. One-to-N & N-to-One: two advanced backdoor attacks against deep learning models[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(3): 1562-1578.

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

[35] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning [EB/OL]. (2017-12-15)[2023-07-12]. https://arxiv.org/abs/1712.05526.

[36] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks [EB/OL]. (2019-12-05) [2023-07-12]. https://arxiv.org/abs/1912.02771.

[37] LI S, XUE M, ZHAO B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[EB/OL]. (2019-09-06)[2023-07-12]. https://arxiv.org/abs/1909.02742v3.

[38] ALGHAZZAWI D M,RABIE O B J, BHATIA S, et al. An improved optimized model for invisible backdoor attack creation using steganography[J]. Computers, Materials & Continua, 2022, 72(1): 1173-1193.

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

[40] ZENG Y, PARK W, MAO Z M, et al. Rethinking the backdoor attacks' triggers: a frequency perspective [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. [S. l.]: IEEE, 2021: 16453-16461.

[41] KWON H, KIM Y. BlindNet backdoor: attack on deep neural network using blind watermark[J]. Multimedia Tools and Applications, 2022, 81(5): 6217-6234.

[42] WANG T, YAO Y, XU F, et al. Backdoor attack through frequency domain [EB/OL]. (2021-11-22) [2023-07-12]. https://arxiv.org/abs/2111.10991v1.

[43] ZHONG N, QIAN Z, ZHANG X. Imperceptible backdoor attack: from input space to feature representation[EB/OL]. (2022-05-06)[2023-07-12]. https://arxiv.org/abs/2205.03190.

[44] LI Y Z, LI Y M, WU B Y, et al. Invisible backdoor attack with sample-specific triggers[EB/OL]. (2020-12-07) [2023-07-12]. https://arxiv.org/abs/2012.03816.

[45] ZHAO B Y, LAO Y J. Towards class-oriented poisoning attacks against neural networks[C]//Proceedingd of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE, 2022: 2244-2253.

[46] HU S, ZHOU Z, ZHANG Y, et al. BadHash: invisible backdoor attacks against deep hashing with clean label[C]//Proceedings of the 30th ACM International Conference on Multimedia. [S. l.: s. n.], 2022: 678-686.

[47] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI,USA: IEEE, 2017: 86-94.

[48] LIAO C, ZHONG H T, SQUICCIARINI A, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[EB/OL]. (2018-08-30)[2023-07-12]. https://arxiv.org/abs/1808.10307.

[49] LIU Y Q, MA S Q, AAFER Y, et al. Trojaning attack on neural networks [C]//Proceedings of 2017 Conference on Network and Distributed System Security Symposium. [S. l.: s. n.],2017.

[50] TURNER A, TSIPRAS D, MADRY A. Clean-label backdoor attacks [C]//Proceedings of 2018 International Conference on Learning Representations. [S. l.: s. n.],2018.

[51] ZHANG Q, DING Y, TIAN Y, et al. AdvDoor: adversarial backdoor attack of deep learning system [C]//Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. Denmark: ACM, 2021: 127-138.

[52] SOURI H, FOWL L, CHELLAPPA R, et al. Sleeper agent: scalable hidden trigger backdoors for neural networks trained from scratch[EB/OL]. (2022-10-13) [2023-07-12]. https://arxiv.org/abs/2106.08970.

[53] SHOKRI R, OTHERS. Bypassing backdoor detection algorithms in deep learning[C]//Proceedings of 2020 IEEE European Symposium on Security and Privacy.

[S. l. : s. n.], 2020: 175-183.

[54] NGUYEN A, TRAN A. WaNet—Imperceptible warping-based backdoor attack[EB/OL]. (2021-03-04)[2023-07-12]. https://arxiv.org/abs/2102.10369v3.

[55] QUIRING E, RIECK K. Backdooring and poisoning neural networks with image-scaling attacks[C]//Proceedings of 2020 IEEE Security and Privacy Workshops. [S. l. : s. n.], 2020: 41-47.

[56] XIAO Q, CHEN Y, SHEN C, et al. Seeing is not believing: camouflage attacks on image scaling algorithms[C]//Proceedings of 2019 USENIX Security Symposium. [S. l. : s. n.],2019: 443-460.

[57] LIU Y, MA X, BAILEY J, et al. Reflection backdoor: a natural backdoor attack on deep neural networks[C]//Proceedings of 2020 European Conference on Computer Vision. [S. l.]: Springer, Cham, 2020: 182-199.

[58] HENDRYCKS D, ZHAO K, BASART S, et al. Natural adversarial examples[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 15257-15266.

[59] LI Y, ZHAI T, JIANG Y, et al. Backdoor attack in the physical world[EB/OL]. (2021-04-06)[2023-07-12]. https://arxiv.org/abs/2104.02361v1.

[60] HE C, XUE M, WANG J, et al. Embedding backdoors as the facial features: invisible backdoor attacks against face recognition systems[C]//Proceedings of ACM Turing Celebration Conference-China. [S. l. : s. n.],2020: 231-235.

[61] XUE M, HE C, WANG J, et al. Backdoors hidden in facial features: a novel invisible backdoor attack against face recognition systems[J]. Peer-to-Peer Networking and Applications, 2021, 14(3): 1458-1474.

[62] SUN Y, WANG X, TANG X. Deep learning face representation from predicting 10,000 classes[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S. l. : s. n.], 2014: 1891-1898.

[63] PARKHI O, VEDALDI A, ZISSERMAN A. Deep face recognition[C]//Proceedings of 2015 British Machine Vision Conference. [S. l. : s. n.], 2015.

[64] SARKAR E, BENKRAOUDA H, MANIATAKOS M. FaceHack: triggering backdoored facial recognition systems using facial characteristics[EB/OL]. (2020-06-20) [2023-07-12]. https://arxiv.org/abs/2006.11623.

[65] GUO W, TONDI B, BARNI M. A master key backdoor for universal impersonation attack against DNN-based face verification[J]. Pattern Recognition Letters, 2021, 144: 61-67.

[66] XUE M, HE C, SUN S, et al. Robust backdoor attacks against deep neural networks in real physical world[J]. Computers & Security, 2022, 118: 102726.

[67] BARNI M, KALLAS K, TONDI B. A new backdoor attack in CNNS by training set corruption without label poisoning[C]//Proceedings of 2019 IEEE International Conference on Image Processing. [S. l. : s. n.], 2019: 101-105.

[68] SAHA A, SUBRAMANYA A, PIRSIAVASH H. Hidden trigger backdoor attacks[C]//Proceedings of AAAI Conference on Artificial Intelligence. [S. l. : s. n.],2020: 11957-11965.

[69] NING R, LI J, XIN C, et al. Invisible poison: a blackbox clean label backdoor attack to deep neural networks[C]//Proceedings of 2021 IEEE Conference on Computer Communications. [S. l.]: IEEE, 2021: 1-10.

[70] AGHAKHANI H, MENG D, WANG Y X, et al. Bullseye polytope: a scalable clean-label poisoning attack with improved transferability[C]//Proceedings of 2021 IEEE European Symposium on Security and Privacy. [S. l. : s. n.],2021: 159-178.

[71] ZENG Y, PAN M, JUST H A, et al. Narcissus: a practical clean-label backdoor attack with limited information[EB/OL]. (2022-04-15)[2023-07-12]. https://arxiv.org/abs/2204.05255v2.

[72] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks[EB/OL]. (2018-04-03)[2023-07-12]. https://arxiv.org/abs/1804.00792.

[73] ZHU C, HUANG W R, LI H, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proceedings of 2019 International Conference on Machine Learning. [S. l. : s. n.],2019: 7614-7623.

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

[75] ZOU M, SHI Y, WANG C, et al. PoTrojan: powerful neural-level Trojan designs in deep learning models[EB/OL]. (2018-02-08)[2023-07-12]. https://arxiv.org/abs/1802.03043v1.

[76] CHENG S, LIU Y, MA S, et al. Deep feature space Trojan attack of neural networks by controlled detoxification[J]. The Association for the Advance of Artificial Intelligence, 2021(7):1148-1156.

[77] ZHU J Y, PARK T, ISOLA P, et al. Unpaired im-

age-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2242-2251.

[78] GARG S, KUMAR A, GOEL V, et al. Can adversarial weight perturbations inject neural backdoors? [C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. [S. l. : s. n.], 2020: 2029-2032.

[79] DUMFORD J, SCHEIRER W. Backdooring convolutional neural networks via targeted weight perturbations[C]//Proceedings of 2020 IEEE International Joint Conference on Biometrics. [S. l. : s. n.], 2020.

[80] GUO C, WU R, WEINBERGER K Q. Trojannet: embedding hidden Trojan horse models in neural networks[EB/OL]. (2020-02-24)[2023-07-12]. https://arxiv.org/abs/2002.10078.

[81] RAKIN A S, HE Z, FAN D. Tbt: targeted neural network attack with bit Trojan[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l. : s. n.], 2020: 13198-13207.

[82] CLEMENTS J, LAO Y. Backdoor attacks on neural network operations[C]//Proceedings of 2018 IEEE Global Conference on Signal and Information Processing. Anaheim, CA, USA: IEEE, 2018: 1154-1158.

[83] YAO Y, LI H, ZHENG H, et al. Regula Sub-rosa: latent backdoor attacks on deep neural networks[EB/OL]. (2019-05-24)[2023-07-12]. https://arxiv.org/abs/1905.10447v1.

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

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

[86] SALEM A, BACKES M, ZHANG Y. Don't trigger me! A triggerless backdoor attack against deep neural networks[EB/OL]. (2020-10-07)[2023-07-12]. https://arxiv.org/abs/2010.03282.

[87] YANG W, LI L, ZHANG Z, et al. Be careful about poisoned word embeddings: exploring the vulnerability of the embedding layers in NLP models[C]//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: human Language Technologies. [S. l. : s. n.], 2021: 2048-2058.

[88] CLEMENTS J, LAO Y. Hardware Trojan attacks on neural networks[EB/OL]. (2018-06-14)[2023-07-12]. http://arxiv.org/abs/1806.05768.

[89] LI W, YU J, NING X, et al. Hu-fu: hardware and software collaborative attack framework against neural networks[C]//Proceedings of 2018 IEEE Computer Society Annual Symposium on VLSI. [S. l. : s. n.], 2018: 482-487.

[90] ZHAO Y, HU X, LI S, et al. Memory Trojan attack on neural network accelerators[C]//Proceedings of 2019 Design, Automation & Test in Europe Conference & Exhibition. Florence, Italy: IEEE, 2019: 1415-1420.

[91] BAGDASARYAN E, SHMATIKOV V. Blind backdoors in deep learning models[C]//Proceedings of the 30th USENIX Security Symposium. [S. l. : s. n.], 2021.

[92] COSTALES R, MAO C, NORWITZ R, et al. Live Trojan attacks on deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA: IEEE, 2020: 3460-3469.

[93] TRAN B, LI J, MA A. Spectral signatures in backdoor attacks[C]//Proceedings of 2018 Conference and Workshop on Neural Information Processing Systems. [S. l. : s. n.], 2018.

[94] CHAN A, ONG Y S. Poison as a cure: detecting & neutralizing variable-sized backdoor attacks in deep neural networks[EB/OL]. (2019-11-19)[2023-07-12]. https://arxiv.org/abs/1911.08040.

[95] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[EB/OL]. (2018-11-9)[2023-07-12]. https://arxiv.org/abs/1811.03728.

[96] PERI N, GUPTA N, HUANG W R, et al. Deep k-NN defense against clean-label data poisoning attacks[C]//Proceedings of Computer Vision-ECCV 2020 Workshops. [S. l. : s. n.], 2020: 55-70.

[97] XIANG Z, MILLER D J, KESIDIS G. A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense[C]//Proceedings of the 29th International Workshop on Machine Learning for Signal Processing. [S. l. : s. n.], 2019: 1-6.

[98] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition. [S. l.]:IEEE, 2018: 8789-8797.

[99] GAO Y S, XU C G, WANG D R, et al. STRIP: a defence against Trojan attacks on deep neural networks [C]//Proceedings of the 35th Annual Computer Security Applications Conference. [S. l. :s. n.],2019: 113-125.

[100] UDESHI S, PENG S, WOO G, et al. Model agnostic defence against backdoor attacks in machine learning[J]. IEEE Transactions on Reliability, 2022, 71(2):880-895.

[101] CHOU E, TRAMER F, PELLEGRINO G. SentiNet: detecting localized universal attacks against deep learning systems[C]//Proceedings of 2020 IEEE Security and Privacy Workshops. San Francisco, CA, USA: IEEE, 2020: 48-54.

[102] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization [C]//Proceedings of 2017 IEEE International Conference on Computer Vision. [S. l. :s. n.],2017: 618-626.

[103] LIU Y, XIE Y, SRIVASTAVA A. Neural Trojans [EB/OL]. (2017-10-03)[2023-07-12]. https://arxiv.org/abs/1710.00942v1.

[104] DOAN B G, ABBASNEJAD E, RANASINGHE D C. Februus: input purification defense against Trojan attacks on deep neural network systems[J]. Annual Computer Security Applications Conference, 2020(8): 897-912.

[105] CHATTOPADHYAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: improved visual explanations for deep convolutional networks [C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. [S. l.]:IEEE,2018:839-847.

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

[107] VILLARREAL-VASQUEZ M, BHARGAVA B. ConFoc: content focus protection against Trojan attacks on neural networks [EB/OL]. (2020-07-01)[2023-07-12]. https://arxiv.org/abs/2007.00711.

[108] ZHU L, NING R, WANG C, et al. GangSweep: sweep out neural backdoors by GAN[C]//Proceedings of the 28th ACM International Conference on Multimedia. [S. l.]:ACM, 2020: 3173-3181.

[109] QIU H, ZENG Y, GUO S, et al. DeepSweep: an evaluation framework for mitigating DNN backdoor attacks using data augmentation[C]//Proceedings of 2021 ACM Asia Conference on Computer and Communications Security. [S. l. :s. n.],2021: 363-377.

[110] HUANG K, LI Y, WU B, et al. Backdoor defense via decoupling the training process[EB/OL]. (2022-02-05)[2023-07-12]. https://arxiv.org/abs/2202.03423v1.

[111] LI Y, LYU X, KOREN N, et al. Anti-backdoor learning: training clean models on poisoned data[J]. Advances in Neural Information Processing Systems, 2021, 34: 14900-14912.

[112] 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.

[113] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Proceedings of 2018 Symposium on Research in Attacks, Intrusions and Defenses. [S. l. :s. n.], 2018.

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

[115] LIU Y, LEE W C, TAO G, et al. ABS: scanning neural networks for back-doors by artificial brain stimulation[C]//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. [S. l.]:ACM, 2019: 1265-1282.

[116] WANG Y, ZHAI D H, HE Y, et al. An adaptive robust defending algorithm against backdoor attacks in federated learning[J]. Future Generation Computer Systems, 2023, 143: 118-131.

[117] ZHU L, NING R, XIN C, et al. CLEAR: clean-up sample-targeted backdoor in neural networks [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. [S. l.]: IEEE, 2021: 16433-16442.

[118] CHEN H, FU C, ZHAO J, et al. DeepInspect: a black-box Trojan detection and mitigation framework for deep neural networks [C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. [S. l. :s. n.],2019: 4658-4664.

[119] KOLOURI S, SAHA A, PIRSIAVASH H, et al. Universal litmus patterns: revealing backdoor attacks in CNNS[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 298-307.

[120] XU X J, WANG Q, LI H C, et al. Detecting AI Trojans using meta neural analysis[C]//Proceedings

of 2021 IEEE Symposium on Security and Privacy. New York:IEEE, 2021:103-120.

[121] VELDANDA A K, LIU K, TAN B, et al. NNoculation: catching badnets in the wild[C]//Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. [S. l.]: ACM, 2021: 49-60.

[122] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization [EB/OL]. (2018-02-28)[2023-07-12]. https://arxiv.org/abs/1711.01991.

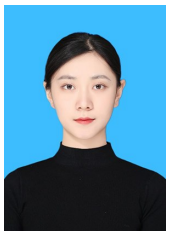
[123] GUO W, WANG L, XING X, et al. TABOR: a highly accurate approach to inspecting and restoring Trojan backdoors in AI systems[EB/OL]. (2019-08-02)[2023-07-12]. https://arxiv.org/abs/1908.01763v1.

[124] LI Y G, LYU X X, KOREN N, et al. Neural attention distillation: erasing backdoor triggers from deep neural networks[EB/OL]. (2021-01-15)[2023-07-12]. https://arxiv.org/abs/2101.05930.

[125] DU M, JIA R, SONG D. Robust anomaly detection and backdoor attack detection via differential privacy [EB/OL]. (2019-11-16)[2023-07-12]. https://arxiv.org/abs/1911.07116v1.

[126] JAVAHERIPI M, SAMRAGH M, FIELDS G, et al. CLEANN: accelerated Trojan shield for embedded neural networks [EB/OL]. (2020-09-04)[2023-07-12]. https://arxiv.org/abs/2009.02326.

作 者 简 介



孙钰媛

女,1996 年生,博士研究生,研究方向为人工智能安全

E-mail:sun_yuyuan@nudt.edu.cn

王 璇

女,1999 年生,博士研究生,研究方向为人工智能安全

E-mail:wangxuan21d@nudt.edu.cn

陆余良

男,1964 年生,博士,教授,博士研究生导师,研究方向为软件与系统安全、网络态势感知、大数据分析

E-mail:luyuliang@nudt.edu.cn

责任编辑 钱 静