计算机科学
COMPUTER SCIENCE

深度学习模型的后门攻击研究综述

应宗浩，吴槟

引用本文

应宗浩，吴槟. 深度学习模型的后门攻击研究综述[J]. 计算机科学, 2023, 50(3): 333-350.

YING Zonghao, WU Bin. Backdoor Attack on Deep Learning Models:A Survey[J]. Computer Science, 2023, 50(3): 333-350.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600031

深度学习模型的后门攻击研究综述

应宗浩　吴　槟
中国科学院信息工程研究所信息安全国家重点实验室　北京 100085
中国科学院大学网络空间安全学院　北京 100049
(yingzonghao@iie.ac.cn)

摘　要　近年来,以深度学习为代表的人工智能在理论与技术上取得了重大进展,在数据、算法、算力的强力支撑下,深度学习受到空前的重视,并被广泛应用于各领域。与此同时,深度学习自身的安全问题也引起了广泛的关注。研究者发现深度学习存在诸多安全隐患,其中在深度学习模型安全方面,研究者对后门攻击这种新的攻击范式进行广泛探索,深度学习模型在全生命周期中都可能面临后门攻击威胁。首先分析了深度学习面临的安全威胁,在此基础上给出后门攻击技术的相关背景及原理,并对与之相近的对抗攻击、数据投毒攻击等攻击范式进行区分。然后对近年来有关后门攻击的研究工作进行总结与分析,根据攻击媒介将攻击方案分为基于数据毒化、基于模型毒化等类型,随后详细介绍了后门攻击针对各类典型任务及学习范式的研究现状,进一步揭示后门攻击对深度学习模型的威胁。随后梳理了将后门攻击特性应用于积极方面的研究工作。最后总结了当前后门攻击领域面临的挑战,并给出未来有待深入研究的方向,旨在为后续研究者进一步推动后门攻击和深度学习安全的发展提供有益参考。

关键词:深度学习;模型安全;后门攻击;攻击范式;数据毒化
中图法分类号　TP391

Backdoor Attack on Deep Learning Models:A Survey

YING Zonghao and WU Bin
State Key Laboratory of Information Security,Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100085,China
School of Cyber Security,University of Chinese Academy of Sciences,Beijing 100049,China

Abstract　In recent years,artificial intelligence represented by deep learning has made breakthroughs in theories and technologies. With the strong support of data,algorithms and computing power,deep learning has received unprecedented attention and has been widely used in various fields,bringing great improvements to the corresponding fields. With the wide application of deep learning technology in various fields including security critical ones,the security issue of deep learning has attracted more and more attention. Researchers have found many security risks in deep learning systems. In terms of the security of deep learning models,researchers have extensively explored the new attack paradigm of backdoor attack. Backdoor attack can threaten deep learning models throughout their whole life cycle. A large number of researchers have proposed series of attack scheme from different angles. This paper takes the security threats of deep learning system as a starting point,introduces the current attack paradigms. On this basis,it gives the back-ground and principle of backdoor attack,distinguishes the similar attack paradigms such as adversarial attack and data poisoning attack,then continues to elaborate on the attack principle and outstanding features of the classic methods of backdoor attack to date. According to the working principle,the attack schemes are divided into data poisoning based attack and model poisoning based attack and others,the paper systematically summarizes them and clarify the advantages and disadvantages of current research. Then,this paper surveys the state-of-the-art works of backdoor attack against various typical applications and popular deep learning paradigms,which further reveal the threat of backdoor attack towards deep learning models. Finally,this paper summarizes the research work on applying backdoor attack characteristics to positive applications and explores the current challenges of backdoor attack,as well as discusses future research directions worthy of in-depth exploration,

到稿日期:2022-06-02　返修日期:2022-11-19
基金项目:国家自然科学基金(U1936119,62272007);海南省重大科技计划(ZDKJ2019003);中国国家铁路集团有限公司科技研究开发计划项目(N2021W003,N2021W004)
This work was supported by the National Natural Science Foundation of China(U1936119,62272007),Major Technology Program of Hainan,China(ZDKJ2019003) and Science and Technology Research and Development Program Project of China State Railway Group Co.,Ltd(N2021W003,N2021W004).
通信作者:吴槟(wubin@iie.ac.cn)

aiming to provide guidance for the follow-up researchers to further promote the development of backdoor attack and security of deep learning.

Keywords Deep learning, Security of model, Backdoor attack, Attack paradigms, Data poisoning

1 引言

自从 2012 年 AlexNet[1] 在 ImageNet 大规模图像识别竞赛中(ImageNet Large Scale Visual Recognition Challenge, ILSVRC)以极大的优势超越传统方案后,深度学习开始得到研究者的重视。此后深度学习的效果不断得到提升,Simonyan 等于 2014 年提出的 VGG[2] 及同年提出的 GoogleNet[3] 将 Top 5 正确率提升到 90%以上,He 等提出的 ResNet[4] 的正确率更是超过了人类的水平。深度学习的应用无处不在,不仅包括计算机视觉领域,还包括自动驾驶[5]、语音识别[6]、情感分析[7]、医学图像分析[8]等。深度学习的应用提升了人类生活质量,但对于安全性敏感的领域(自动驾驶[9]、垃圾邮件过滤[10]、欺诈检测[11])而言,如果深度学习遭受攻击,将造成巨大的损失。以自动驾驶系统为例,BadNets[12] 成功攻击了自动驾驶系统,使其将"停止"标志识别为"限速"标志,而在 Liu 等[13] 设计的 TrojanNN 中,自动驾驶系统在受到攻击后会偏离正确的行驶方向。基于深度学习的应用无处不在,研究者需及早考虑深度学习的安全问题。本文将深度学习的攻击面依据其生命周期进行划分,如图 1 所示,其中模型逆向攻击(Model Inversion Attack)、模型提取攻击(Model Extraction Attack)属于隐私威胁,它们会破坏系统的机密性,而数据投毒攻击(Data Poisoning Attack)、后门攻击(Backdoor Attack)、对抗攻击(Adversarial Attack)则属于完整性威胁和可用性威胁,它们会降低模型的测试性能。在数据收集阶段,深度学习系统可能面临数据投毒攻击[14-16],攻击通过篡改训练集中部分样本的内容和标签来实现。在模型测试阶段会受到对抗攻击[17-19]、模型逆向攻击[20-22]和模型窃取攻击[23-25]。作为攻击者,在对抗攻击中只需对输入数据做出轻微扰动即可欺骗模型做出错误决策,在模型逆向攻击中可以从模型预测结果中提取与训练数据有关的敏感信息,在模型窃取攻击中通过对原模型进行一定次数的查询,就可以窃取模型参数进而重构原模型。

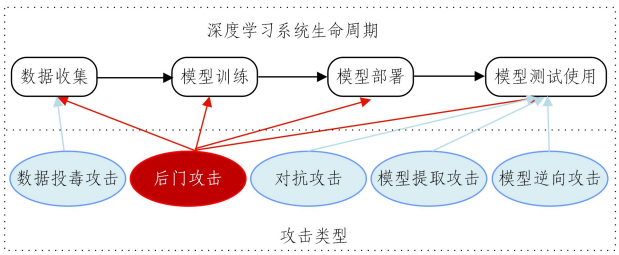


图 1 深度学习攻击面

Fig. 1 Attack scenarios for deep learning

这些攻击手段仅涉及深度学习生命周期的某一阶段,而后门攻击可以存在于深度学习全生命周期。后门攻击分为两个阶段,分别是植入后门和触发后门。植入后门可能发生在生命周期的任一阶段,而触发后门则发生在模型测试阶段。由于深度学习模型本质上是一组与特定结构相连接的权重矩阵,其定义了输入和输出之间复杂的非线性关系,因此解释模型的决策是非常困难的[26-28],同时模型在受到后门攻击后不会显著降低在主任务上的表现,因此后门攻击具有极大的隐蔽性。

本文综述了深度学习模型后门攻击的最新研究进展和研究方向,主要贡献如下:

(1)系统梳理并分析了深度学习后门攻击的典型方案,依据攻击媒介对后门攻击技术进行具体划分,对每种类型的攻击技术进行剖析,比较了不同技术之间的联系与优劣。

(2)调研了后门攻击在具体场景下的应用。在消极应用方面,后门攻击会给包括计算机视觉、自然语言处理在内的应用带来安全风险;在积极应用方面,后门攻击能够促进模型、数据所有权的保护等。这进一步明确了后门攻击的研究对基于深度学习的应用的重要性。

(3)分析了后门攻击对包括迁移学习、联邦学习、图神经网络等学习范式造成的威胁,指出了后门攻击的研究对深度学习领域发展的重要意义。

(4)探讨了后门攻击面临的主要挑战,并给出未来值得进一步探索的发展方向。

本文第 1 节介绍了深度学习及其面临的安全风险并由此引入了后门攻击;第 2 节对后门攻击的背景进行阐述,包括术语、定义、攻击场景以及与其他攻击范式的联系;第 3 节以不同攻击媒介对后门攻击方法进行归纳与分析,并给出评估指标及基础研究工作进展;第 4 节概述了面临后门攻击的典型应用;第 5 节介绍了面临后门攻击的其他学习范式;第 6 节介绍了后门攻击的积极应用;第 7 节从触发器、评估指标、可解释性研究等多方面进行分析与展望;最后对当前后门攻击领域的研究及本文贡献进行总述。

2 背景

后门攻击是针对深度学习模型的新攻击范式,本节介绍相关概念、攻击场景以及与其他攻击方案的联系。

2.1 后门攻击术语

后门攻击作为一种的新的攻击范式,涉及大量领域术语,为便于表述,表 1 列出了与后门攻击相关的术语。

表 1 相关术语

Table 1 Related terms

术语	符号	描述
触发器	Δ	用于构造毒化样本,在测试时能触发后门
目标原样本	x_t	属于目标类的原样本
目标标签	y_t	攻击者指定标签,攻击成功则毒化样本被预测为 y_t
其他原样本	x_{nt}	属于非目标类的原样本
其他标签	y_{nt}	目标标签以外的标签,是 x_{nt} 对应的标签
原样本	x	包括目标原样本 x_t 和其他原样本 x_{nt}
原标签	y	原样本对应的标签
样本毒化函数	$U(x,\Delta)$	输入原样本和触发器,输出为毒化样本
标签毒化函数	$V(y)$	输入原标签,输出依据攻击类型而异,$y_t=V(y)$
毒化样本	x'	原样本上注入触发器得到毒化样本,$x'=U(x,\Delta)$
原模型	F_ω	攻击者的目标模型,ω 为原模型权重
后门模型	F_{ω^*}	被植入后门的模型,ω^* 为后门模型权重

2.2 后门攻击定义

后门攻击原指在系统中隐藏恶意功能的技术，它只能由某个条件触发。深度学习中的后门攻击指将后门植入原模型中，在测试阶段，若样本中存在触发器则能够触发后门，模型由此得出特定的预测结果，否则模型表现正常。深度学习后门攻击的形式化描述如下：

在植入后门阶段，给定原模型 F_ω（ω 表示原模型权重），攻击者通过数据毒化、模型毒化等方式，根据攻击方案特点，设计有效的样本毒化函数 $U()$、触发器 Δ 等，将后门植入原模型中得到后门模型 F_{ω^*}（ω^* 表示后门模型权重）。在触发后门阶段，后门模型依据输入样本中是否存在触发器而分别执行主任务或后门任务，表示如下：

$$\begin{cases} F_{\omega^*}(x)=F_\omega(x)=y \\ F_{\omega^*}(x')=y_t \end{cases} \tag{1}$$

其中，$x'=U(x,\Delta)$ 为毒化样本，y 为原标签，y_t 为目标标签。

主任务要求后门模型对原样本的预测与原模型对原样本的预测相同，都能预测正确，以体现后门攻击的隐蔽性；后门任务要求后门模型将毒化样本分类到目标类，以体现后门攻击的有效性。以图 2 所示图像识别任务为例，图 2(a)中的模型是原模型，模型执行主任务，在原样本输入时做出正常预测，而毒化样本输入时模型的预测不会受到触发器的影响；图 2(b)中的模型是后门模型，原样本输入时模型执行主任务，将其预测为飞机，而由于毒化样本右下角带有触发器，因此模型执行后门任务，将其预测为目标类汽车。

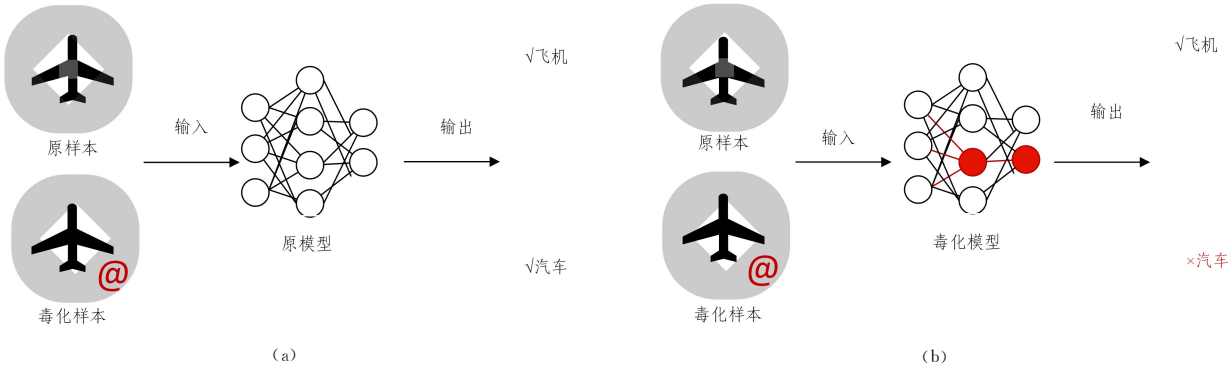


图 2 后门攻击示意图

Fig. 2 Illustration of backdoor attack

2.3 后门攻击场景

后门攻击适用的攻击场景多样，依据攻击媒介划分，攻击者可以通过控制深度学习供应链源头，给用户提供毒化后的第三方代码、硬件、数据等组件，使模型训练过程受到毒化组件的影响从而被植入后门。如果攻击者能够控制训练平台，那么其可直接操纵模型而无须借助毒化数据等媒介。在模型部署甚至测试阶段，攻击者通过修改内存数据等方式依然能够实现后门攻击。

图 3 给出了深度学习生命周期中利用不同攻击媒介可能导致的后门攻击场景。

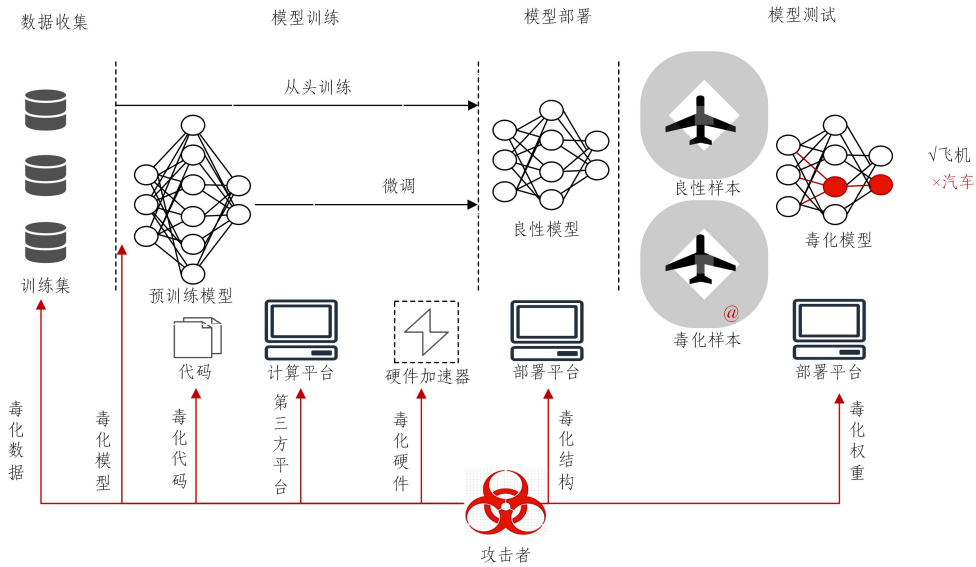


图 3 后门攻击场景

Fig. 3 Attack scenarios of backdoor attack

2.4 与其他攻击方案的联系

后门攻击与对抗攻击、数据投毒攻击都会导致深度学习模型在测试阶段做出错误决策。本小节分析后门攻击与对抗攻击、数据投毒攻击的联系。

2.4.1 与对抗攻击的联系

对抗攻击指使用对抗样本(在原始样本中加入人眼难以感知的对抗扰动所生成的样本)欺骗模型做出错误决策的攻击范式。对抗攻击与后门攻击的联系主要表现在攻击目的、攻击阶段和攻击机制上。

(1)攻击目的。对抗攻击与后门攻击都能降低模型在特定样本上的准确性,对抗攻击中使用的测试样本被称为对抗样本,通过在原样本上叠加对抗扰动得到,其是未知的,需要通过特定优化过程在攻击时实时生成,特定于原样本和模型,通常是不规则的模式;后门攻击使用的样本被称为毒化样本,通过在原样本上注入触发器得到,且该触发器与原模型和原样本无关,应用在任一样本上都能生效,触发器模式可以由攻击者指定,具有最大的灵活性和通用性。

(2)攻击阶段。对抗攻击仅需在模型测试阶段构造对抗样本即可进行攻击;而后门攻击包括植入与触发两个阶段,攻击者可能在任意阶段植入后门并在测试阶段触发后门。

(3)攻击机制。对抗扰动能生效是利用了模型的固有缺陷和样本自身的特点,而触发器能生效是利用了模型学习到的触发器与目标类之间的联系。因此对抗攻击具有迁移性,可以攻击黑盒模型,而后门攻击只能对已植入后门的模型发动攻击。对抗攻击中还存在一种特殊的攻击方式,这种方式被称为通用对抗攻击[29-30],其所用的对抗扰动是通用的,即每个样本使用的对抗扰动都是一样的,这与后门攻击中的触发器非常相似,两者之间的区别也是在于攻击机制的不同。

2.4.2 与数据投毒攻击的联系

数据投毒攻击通过修改模型训练集以降低模型在测试阶段的准确率。根据攻击目标不同,可以分为非定向投毒攻击与定向投毒攻击,前者旨在全面降低模型性能,后者与后门攻击的目标一致。

后门攻击与数据投毒攻击都是为了破坏模型,前者的核心在于向模型中植入后门,攻击方法较多,可以通过直接修改模型结构、权重等实现,数据毒化只是植入后门的途径之一;后者则只能通过数据投毒的方法实现。

在后门攻击中,基于数据毒化的后门攻击与数据投毒攻击较为相似,两者均需要在数据收集阶段进行投毒,但它们的目的不同:后者是为了全面降低模型的性能,而前者要求后门模型测试原样本时的准确率尽可能维持不变,只有当样本中存在触发器时,模型的输出才会出错。相比之下,后门攻击更加隐蔽。而与定向投毒攻击相比,后门攻击需要借助触发器,任一样本中存在触发器才能实现攻击,而前者不需要对测试样本进行修改即可实现攻击,这也意味着只有特定的一部分样本才满足条件,攻击范围相比后门攻击更受限。

3 后门攻击方法

本节依据后门攻击媒介(数据、代码、硬件等)对本领域近年来的代表性研究成果进行归纳分析。后门攻击在早期通过数据毒化实现,此后朝两个方向发展,分别是继续在数据毒化方向上深入研究,提出新的触发器设计方案,以及研究如何通过其他方式植入并触发后门。

3.1 基于数据毒化

基于数据毒化的攻击方案指通过篡改训练集的方式实现后门攻击,这是后门攻击领域被研究得最透彻的攻击类型。不失一般性,基于数据毒化植入后门的方案可以形式化为如下两种类型:

(1)从头训练得到后门模型

给定原数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$,其中 x_i 是样本,y_i 是对应标签。攻击者将原数据集划分为两个子集,分别是待毒化数据集 $D_c=\{(x_i,y_i)\}_{i=1}^{m}$ 和原数据子集 $D_o=\{(x_i,y_i)\}_{i=m+1}^{n}$,其中 m 为毒化样本数。对于每个样本 $(x_i,y_i)\in D_p$,应用样本毒化函数 $U()$ 和标签毒化函数 $V()$ 得到毒化数据集 $D_p=\{(x',y_t)\}_{i=1}^{m}$,其中 $x'=U(x,\Delta)$,$y_t=V(y)$。最后在总数据集 D_t 上进行训练得到后门模型 F_{w^*},$D_t=D_p\cup D_o$。

(2)重训练原模型得到后门模型

给定目标模型 F_w,攻击者构造毒化数据集 $D_p=\{(x',y_t)\}_{i=1}^{m}$,用其重训练 F_w 后得到后门模型 F_{w^*}。

根据构造 D_p 时是否需要修改毒化样本的标签,攻击被划分为毒化标签攻击以及清洁标签攻击,前者需要修改毒化样本的标签为目标标签,后者则不需要。

在模型测试阶段,模型能根据样本中是否有触发器而选择执行主任务或后门任务,如式(1)所示。

3.1.1 毒化标签攻击

根据攻击特点,毒化标签攻击可以分为 3 类。第一类是基础攻击,其特点是可率先实现后门攻击,并初步提升攻击的隐蔽性,但是这类方案采用的触发器容易被察觉。为了提高隐蔽性,研究者提出了两种设计思路,即隐蔽攻击(Invisible Attack)和不可感知攻击(Imperceptible Attack),前者的目的在于隐藏触发器,而后者希望通过精心设计的触发器模式,确保触发器不会被人类感知为异常。

(1)基础攻击

Gu 等提出的 BadNets[12] 首次在深度学习模型中实现了后门攻击。其攻击方式遵循前文形式化后的从头训练得到后门模型,这种攻击方式被后续的很多工作借鉴并加以改进。BadNets 的提出对本领域具有重大的研究意义,它不仅表明后门攻击在深度学习中是可行的,同时指出了后门攻击对深度学习供应链安全的重要影响。

文献[12]的工作中要求攻击者可以修改模型原训练集,TrojanNN[13] 削弱了这一假设,生成了能够最大程度激活模型中与目标类相关的神经元的触发器,并通过逆向工程得到训练集,由此进一步生成毒化训练集,最后通过重训练即可植入后门。Chen 等[31] 在一个更弱的威胁假设下进行研究,攻击者甚至能在不知道具体模型的情况下通过数据毒化进行攻击。其研究表明,只需注入 50 个毒化样本就可以达到 90% 以上的成功率。

为了提升攻击的隐蔽性,研究者针对修改样本标签及毒化样本生成两方面进行改进。Peng 等[32] 提出了 label-smoothed attack,毒化样本的标签并不会全被修改为目标标签,而是以概率 $p_n(x)$ 进行修改,$p_n(x)$ 的值控制在确保毒化样本被预测为目标类的概率稍大于其他类的概率,以确保能在攻击成功的同时提升隐蔽性。Ali 等[33] 提出的 low-confi-

dence attack 也做了类似处理。

针对静态（固定模式和位置）触发器容易被当前的防御方法[34-37]检测到的现象，文献[38-39]提出了动态生成触发器的方案。Salem 等[38]通过 3 种不同的技术生成具有随机模式和位置的触发器，包括利用生成对抗网络（Generative Adversarial Network，GAN）[39]进行联合训练，生成最佳的触发器，但本质上还是依赖于统一的触发器，只是将单一的触发器扩展到一组位置、模式的触发器集合。Nguyen 等[40]在此基础上进行扩展，构建以多样性损失（Diversity Loss）驱动的触发器生成网络，确保各样本都有唯一的触发器，并且该触发器对其他样本无效。该方案实现了触发器的不可重用，并成功规避了防御方案[34-35,41]的检测。

以上工作都在图像领域中进行，文本领域的后门攻击机理与其类似，但由于前者是连续（Continuous）的，后者是离散（Discrete）的，并且文本触发器更容易改变样本的语义等信息，因此两者在触发器选择等方面存在差异。图像领域的研究领先于文本领域，后者由此受到许多启发。以文本中的字符、词、语句来类比图像中的像素点、像素组成的模式、图像，可以发现明显的研究规律。BadNets[12]选择扰动像素点或者像素组成的特定模式作为触发器，相应地，在文本领域可以选择单词[42-45]或词序列[13]作为触发器。为了避免误触发，一般选择特殊的词或者词序列，而 Zhang 等[46]使用任意词的逻辑组合作为触发器，使触发器的设计更具灵活性。Chen 等[31]使用了与原样本相同规格的图像作为触发器，调整透明度后叠加于原样本上，相应地，在文本领域也可以直接使用句子作为触发器[43-44,47]。实际上，Chen 等[43]提出了 3 种构造触发器的方法，包括字符（Character）级别、词（Word）级别和语句（Sentence）级别的触发器。其研究表明使用字符级触发器可能导致语义异常，而使用语句级触发器通过转换时态可以维持语义并避免语法检查异常。在攻击效率上，词级触发器优于字符级触发器和语句级触发器。这表明进行攻击时需要在攻击有效性和隐蔽性之间进行权衡。

（2）隐蔽攻击

隐蔽攻击是通过隐藏毒化样本中的触发器来提升后门攻击的隐蔽性，其挑战在于如何在后门攻击的有效性和触发器的隐蔽性之间取得平衡，对此，研究者借鉴了其他领域（对抗样本[48-51]、信息隐藏[52-55]、图形处理[56-61]）的技术。

Liao 等[48]利用扰动掩码（Perturbation Mask）作为触发器，在 Deepfool[50]方案基础上进行改进，能够基于当前样本和模型，找到小幅度的自适应扰动，将选定的样本“推向”目标类的决策边界。Zhang 等[49]分别将 DeepFool[50]和 C&W[51]方案生成的对抗扰动合成为通用扰动，并将其作为触发器，由于利用了数据集的信息，因此毒化样本更接近决策边界，同时也实现了视觉上的隐蔽性。

Li 等[52]分别通过隐写术和正则化的方法来实现触发器的隐藏，前者将最低有效位（Least Significant Bit，LSB）隐写[56]应用于样本毒化函数来实现隐藏，而后者实际是 TrojanNN[13]的改进，它通过优化过程生成触发器，使用范数正则化确保触发器的隐蔽。Xue 等[53]利用离散余弦变换（Discrete Consine Transform，DCT）隐写[57]，在图像的不同通道上加入触发器，从而实现触发器的隐藏。Li 等[54]受到基于 DNN 的隐写术[58-59]的启发，通过编码器-解码器网络将攻击者指定的字符串编码后作为隐形加性噪声，并以此作为触发器，模型训练期间学习字符串与目标标签的映射。Zhang 等[55]选定图像边缘结构作为注入触发器的区域，并在其中嵌入颜色值以生成触发器，然后利用深度注入网络将其嵌入到载体图像中。由于边缘结构属于图像的高频分量，将触发器隐藏其中难以被发现。

Quiring 等[60]利用图像缩放（Image Scaling）技术，当图像被缩放到特定分辨率时其内容会改变，从而实现触发器的隐藏。Wang 等[61]将图像从 RGB 通道转换到 YUV 通道，然后将原始样本分割成不相连的块，并在每个块的 UV 通道的中高频部分注入触发器，这本质上是利用了人类视觉系统对 UV 通道的不敏感性。

文本领域的基础后门攻击都是基于插入的方法，它们可能会造成原样本的语法错误和不流畅，而且无法实现隐蔽。Qi 等[62]使用句法结构作为触发器以避免出现语法错误，这是一个更加抽象和潜在的特征，不容易被感知。Li 等[63]通过同形异义词和动态语句生成来解决不流畅的问题。接受原始 Unicode 字符作为输入的模型利用基于视觉欺骗的 Unicode 同形异义词的字符级触发器来保证毒化样本与原始文本相同的可读性，而动态语句生成技术则利用语言模型生成的高度自然流畅的语句作为后门触发器，两类技术都可以实现触发器的隐藏。

（3）不可感知攻击

不可感知攻击的目的不在于隐藏触发器，而是确保触发器的存在不会让毒化样本变得突兀，即使毒化样本中的触发器非常显眼，人类也不会察觉异常。

文献[31，64-66]通过设计与攻击场景语义相关的触发器来实现不可感知攻击。以攻击人脸识别模型为例，Chen 等[31]选择眼镜和太阳镜作为触发器，但其受限于攻击场景，如果不允许佩戴眼镜则无法触发。He 等[64]将触发器的形状设计为眉毛和胡须的轮廓，在毒化样本时用该触发器替换人脸对应的部位即可，而在测试阶段，攻击者通过化妆即可触发后门。Sarkar 等[65]利用面部动作（微笑、皱眉等）作为触发器，它们可以通过面部肌肉的运动被自然引入，避免了攻击场景的限制。Lin 等[66]设计的方案也是使用语义信息作为触发器，不同点在于，其组合了多个非目标类的原样本的特征作为触发器，当选定的特征组合存在于测试样本中时，则会被识别为目标类别。

除了使用语义相关触发器以外，通过设计人类无法理解或察觉的触发器也可实现不可感知攻击。

Nguyen 等[67]借助图像扭曲（Image Warping）技术，通过扭曲图像原像素生成毒化样本。Doan 等[68]设计了可以注入不可见噪声生成毒化样本的样本毒化函数，同时在相应的经验分布上基于 Wasserstein 的正则化匹配原样本和毒化样本的隐表示（Latent Representation），以确保毒化样本在输入空间和隐空间（Latent Space）中都是难以察觉的。Cheng 等[69]对模型植入后门和清除后门的过程进行迭代，在清除后门期间重训练后门模型，不再对简单的触发器特征敏感，而在下一

轮植入后门期间控制样本毒化函数使用更复杂的触发器特征，最终后门模型学到的是轻微但复杂的触发器特征，人类对此无法理解，但模型能轻易识别出来。

这些工作都是遵循先生成毒化样本，然后训练模型的二阶流程。Doan 等[70]统一两者进行联合学习，将其表述为一个非凸、有约束的优化问题，通过二阶优化过程来解决，从而生成不可感知的、动态的触发器。

文本领域的后门攻击存在容易被感知为异常（因为插入特殊词）或误触发（所插入句子的子集也可触发后门）的情况，Yang 等[71]使用负数据增强来解决误触发问题，且在训练时，通过只更新触发词的嵌入（Embedding）来解决异常问题。Qi 等[72]则使用原文本单词的同义词作为触发器，从而保留原始语义，不再被感知异常。Chan 等[73]通过条件对抗性正则化自动编码器（Conditional Adversarially Regularized Autoencoder，CARA）在文本中植入触发器，以确保毒化样本仍能保持连贯性和语法正确性。

从方案特点来看，基础攻击类型中的方案率先实现了后门攻击，并初步尝试通过改进标签毒化机制[32-33]、触发器生成机制[38-39]来提升数据毒化过程的隐蔽性。隐蔽攻击和不可感知攻击同样都是为了弥补基础攻击的缺陷而提出的，即毒化样本与原样本具有显著差异而容易被检测出来。但两类方案应用的思路不同，因此采用的技术手段也存在差异。隐蔽攻击旨在隐藏触发器，因此会借用对抗攻击、信息隐藏、图像处理等领域的技术，但是其限制了触发器的灵活性；而不可感知攻击旨在确保毒化样本自然，不会因为存在触发器而使人觉得突兀，无须隐藏触发器。为此，图像领域的研究重点在于如何利用语义信息作为触发器[64-66]，而文本领域的研究重点则在于如何保留语义的连贯性和语法的正确性[71-73]。这些工作保留了触发器选择的灵活性，但容易被误触发，因而提升了后门被发现的概率。

3.1.2 清洁标签攻击

基于毒化标签的后门攻击的缺点显而易见：用于训练的毒化样本的标签被修改为目标标签，与毒化样本的原标签差异较大。不论是通过人类的视觉检测还是预分类步骤都能轻易检测出毒化样本。清洁标签攻击指不需要修改毒化样本标签的后门攻击类型，其需要满足两个条件：1）在毒化样本中触发器尽量保持隐蔽；2）毒化样本与其对应的标签在特征上是一致的。

Barni 等[74]通过与 BadNets[12]等工作相反的毒化数据思路实现了清洁标签攻击。攻击者选定目标类 y_t 而不是非目标类 y_{nt} 的一部分样本，应用样本毒化函数 $U(x,\Delta)$ 得到毒化样本 x_t'，不修改标签，此时的样本标签对为 (x_t', y_t)，然后利用此毒化数据集训练模型。该方案的缺点在于需要针对具体的任务和目标类设计不同的 $U(x,\Delta)$，以保证触发器在能够被模型作为目标类的特征进行学习的同时兼具隐蔽性。Barni 等针对手写数字识别和交通标志识别任务分别设计斜坡信号（Ramp Signal）以及水平正弦信号（Horizontal Sinusoidal Signal）作为触发器。Gan 等[75]也通过相似的思路实现文本领域的清洁标签攻击。

由文献[74]的工作可见，后门攻击的关键在于强制模型学习触发器与目标类的联系，只要将属于目标类的样本作为原样本，在其上植入隐蔽的触发器就可以实现清洁标签攻击。研究者只需考虑如何实现触发器的隐蔽。

Liu 等[76]利用反射现象，将反射图案作为触发器进行叠加以实现触发器的隐藏；文献[77-79]通过特征碰撞（Feature Collision）的方法实现攻击，它们的区别在于：文献[77]利用与触发器特征相近的人眼不可见噪声作为触发器；文献[78]利用与非目标类样本 x_{nt} 特征相近的样本作为毒化样本；文献[79]利用与中间样本特征相近的样本作为毒化样本。

Turner 等[80]指出，如果模型在训练过程中学习触发器的模式而不是图像的原始内容，就能更容易学习到触发器与目标类的映射关系，为此分别借助 GAN 和对抗扰动技术先生成更难被学习的中间样本，在其基础上生成毒化样本。Zhao 等将这种攻击扩展至视频识别应用[81]。

根据方案特点结合对应攻击原理可知，文献[78-79]设计的方案通用性较差，主要原因在于毒化图像是通过与其他图像在特征空间碰撞得到，且计算开销较大；文献[74-75]的方案通用性一般，其原理是将触发器与毒化标签相关联，但是其触发器的设计策略根据数据集的不同而不同；文献[76-77，80-81]的方案通用性较好，其原理也是将触发器与毒化标签相关联，文献[76]选择反射映像作为触发器，文献[77]使用不可见噪声作为触发器，而文献[80-81]通过降低模型学习中间样本与目标标签的关联能力，从而强制模型学习触发器与目标标签之间的联系，触发器的设计策略不因数据集而异。

3.2 基于模型毒化

基于数据毒化的方案通过训练来改变模型的权重，因此它们无法被应用于权重被优化、冻结或者不可再训练的模型。此外，当攻击者不具备训练能力时，此类方案同样不适用。为此研究者提出了基于模型毒化的方案，通过直接修改模型的权重或者结构实现后门攻击。

3.2.1 权重毒化攻击

权重毒化的目的是通过调整模型的权重以拟合原模型在毒化数据集上训练后的效果，从而避免模型训练的过程并减小因毒化样本而被检测到攻击行为的概率。

Dumford 等[82]将权重调整的任务转换为对模型的搜索任务，被搜索的模型通过对原模型的权重应用随机加性扰动得到，搜索的目的是找到一个最优的后门模型。为了确保权重扰动对原模型的影响足够小从而实现隐蔽，Hong 等[83]通过消融分析识别出特定的神经元集合，并在每层中选择当原样本和毒化样本输入时具有最大激活差异的神经元，修改其权重，将它们连接到模型的最终输出，实现后门植入。Ji 等[84]通过显著性度量技术确定所需扰动的权重并选择对后门任务影响大、对主任务影响小的参数用于扰动。

Garg 等[85]将对抗扰动应用于模型权重，通过优化的方式得到所需扰动。他们在原模型权重附近进行扰动，使用训练集上的交叉熵损失训练模型，通过优化得到合适的权重。Zhang 等[86]为其优化过程提供了理论解释。Costales 等[87]通过基础后门攻击对原模型进行重训练得到需要修改的权重的新值，然后用新的权重值替换载入到内存中的模型权重值即可。Qi 等的方案[88]与此类似，相较于 Costales 等修改整个

网络的部分权重，Qi 等修改了特定子网的全部权重，他们使用预先设计的恶意子网权重值直接替换原模型中对应的原子网权重值，从而在模型中植入后门。

文献[89-90]通过 Row Hammer[91] 等技术翻转部分权重的比特位来实现后门攻击，其关键在于如何找到待翻转的比特位。Rakin 等[89] 设计了专门的触发器，用于定位存储在 DRAM 中的一部分易受攻击的权重比特，翻转这些比特即可实现攻击。Chen 等[90] 则以贪婪的方式逐步确定模型最脆弱的权重比特位，在每次迭代时找到对后门攻击最敏感的权重元素并针对这个元素定位最适合的比特位用于比特翻转攻击(Bit Flip Attack)。

基于权重毒化实现后门攻击的方案有各自明显的特点，权重扰动方式的不同决定了生成后门模型的方式和攻击目标的不同。就权重扰动方式而言，Dumford 等[82] 通过随机扰动权重得到候选模型并选择其中符合要求的模型作为后门模型，其方案是启发式的，无法确保找到的模型最优，而其余方案的权重均是通过优化或者训练得到的。此外，在生成后门模型时，文献[82，87-90]的方案要求攻击者具有系统安全领域的能力，能够获得模型部署平台的权限或者在其上实现任意代码执行、线程注入等攻击的能力，以实现最终的权重替换。

3.2.2 结构毒化攻击

深度学习模型不仅包括权重，还包括内部的结构，攻击者可以通过修改神经元连接、计算操作、子网等方式植入后门。

Salem 等[92] 基于 dropout 技术实现后门攻击。他们对与目标类关联的神经元所在层应用 dropout 进行训练。通过训练将被 dropout 的神经元和目标类关联。在测试阶段，一旦攻击者应用 dropout，模型就会输出目标标签，否则表现正常。

Clements 等[93] 根据选定的目标操作所在层将网络划分为两个子网，然后基于第一个子网的输出和目标类利用基于 JSMA[94] 改进的算法对目标操作进行特定扰动，最后将两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。

Tang 等提出的 TrojanNet[95] 首先通过基于数据毒化的方式得到毒化子网，然后将其与原模型通过合并层结合得到后门模型，后门模型会合并原模型和毒化子网的输出并进行最终预测。Li 等设计的 DeepPayload[96] 同样是通过在原模型中插入一个恶意子网来实现攻击，首先将模型二进制文件反汇编为数据流图，在其中插入一条旁路，该旁路由触发器检测器和条件模块组成，当触发器检测器检测触发器时，条件模块选择用目标标签替换原模型的原输出。最后重新编译修改后的数据流图得到后门模型。Yang 等[97] 的方案不需要新增子网，仅修改嵌入层(Embedding Layer)即可在语言模型中植入后门。

相较而言，在攻击方案通用性方面，文献[92-93，97]的方案只能用于特定模型，而文献[95-96]的方案可以应用于任意模型，只需在攻击时将其拼接到目标模型即可。两者的区别在于，TrojanNet[95] 是合并结构，后门模型的输出由目标模型和毒化子网共同决定，只是后者权重更大；而 DeePayload[96] 是选择结构，检测到毒化样本后目标模型剩余部分不再被激活，模型输出完全由毒化子网决定。在攻击开销与效率方面，由于 Yang 等[97] 的方案无须训练，因此其效率最高；文献[95-96]训练毒化子网，在攻击时与目标模型拼接实现间接后门植入，而 Clements 等[93] 需要计算雅克比矩阵，较为耗时，它们的攻击效率次之；Salem 等[92] 训练目标模型植入后门，需要参与目标模型的完整训练过程，因此效率较低且开销较大。

3.3 其他方法

基于数据毒化和基于模型毒化的后门攻击分别从数据集和模型的角度考虑。除此之外，作为深度学习系统的重要组成部分，硬件、代码等也有可能被应用于植入后门。

3.3.1 硬件毒化攻击

当前半导体行业中集成电路设计外包、制造全球化、采用第三方 IP 核(Intellectual Property Core)等现象已十分普遍，位于供应链上游的攻击者可以在硬件中植入后门，从而对深度学习模型发动后门攻击。此外，用户可能直接向企业购买训练完成的模型。对于企业而言，为了加速模型计算以及保护知识产权，模型可能是以专用集成电路等形式交付，此时若在硬件层面植入后门，则用户难以检测。

Clements 等[98] 使用与文献[93]相同的方案来确定要扰动的目标操作所在的层及相应扰动，然后基于扰动设计载荷电路，使用数据选择器(MUX)选择被修改后的毒化激活函数从而控制模型的输出。Li 等[99] 提出了软硬件协同的攻击方案，在软件层面将后门植入模型的部分子网，在硬件层面上将恶意电路插入硬件处理单元中。模型运行时在乘法操作完成后会判断后门是否被触发，若被触发，则使用 MUX 选择恶意子网的权重以实现部分加法操作，即仅激活毒化子网，从而控制整个模型的输出。该方案需要毒化全部硬件设施且需要训练原模型，而 Zhao 等[100] 的方案仅通过在内存控制器中植入硬件后门即可实现攻击。神经网络加速器和 DRAM 之间传输的数据会通过内存控制器，它们会对内存访问模式进行监视，识别输入模型的图像数据，并比较图像的频谱以判断是否存在触发器，若存在，则使用错误数据替换正在写入内存的特征图(Feature Map)。

在攻击者假设方面，文献[98-99]要求攻击者能够同时修改硬件电路和模型。为了修改模型，Clements 等[98] 仅通过测试模型表现决定如何修改；Li 等[99] 的方案则需要控制模型的训练过程；Zhao 等[100] 的方案仅要求攻击者对内存控制器进行毒化，当其被神经网络加速器作为第三方 IP 集成后就能发动攻击。在触发器方面，Li 等[99] 基于电气方案设计硬件部件并将其作为触发器，缺点在于触发条件很难精确控制，而优点在于其不需修改图像即可实现攻击。文献[98-99]选择图像作为触发器，其中 Li 等[99] 以特制的具有强鲁棒性的图案作为触发器，在攻击过程中过于明显，因而降低了攻击的隐蔽性。

3.3.2 代码毒化攻击

用户在开发深度学习应用时往往会使用第三方代码，这些代码包括深度学习框架或者他人开源的模型源码等，用户可以直接应用，或者在其基础上自行修改优化。这种情况适应于另一种攻击场景，即代码毒化，攻击者通过修改用户的所用代码仓库中的关键代码来实现后门攻击。

Bagdasaryan 等[101] 通过修改损失函数等关键代码来实现

后门攻击。他们设计了如下损失函数：

$$\mathbb{L}_{\text{total}}=\alpha_1\ \mathbb{L}_m+\alpha_2\ \mathbb{L}_b(+\alpha_3\ \mathbb{L}_e) \tag{2}$$

其中，$\mathbb{L}_m=\mathbb{L}(F_\omega,x,y)$为主任务的损失函数，$\mathbb{L}_b=\mathbb{L}(F_\omega,x',y_t)$为后门任务的损失函数，$\mathbb{L}_e$为可选的用于规避防御措施的损失函数。

通过多重梯度下降算法(Multiple Gradient Descent Algorithm，MGDA)[102]以及 Franke-Wolfe 优化器[103]计算得到合适的系数，最终得到最佳的总损失函数。攻击者在代码库中添加样本毒化函数 $\mathcal{U}$ 和标签毒化函数 $\mathcal{V}$，并使用$\mathbb{L}_{\text{total}}$替换 L_m 即可实现后门攻击。

该方案利用了当前后门防御领域和代码安全领域的盲点，因为毒化后的代码不影响模型架构、权重等后门防御关注的重点，并且毒化后的代码也不会引发缓冲区溢出漏洞、格式化字符串漏洞等代码安全关注的重点。

3.3.3 样本替换攻击

模型权重的更新取决于梯度，梯度具体的数值取决于训练时所用的样本。如果可以找到与毒化样本在训练时产生的梯度相近的原样本，那么使用其对模型进行训练即可实现后门攻击。

Shumailov 等[104]利用此原理，仅通过替换训练样本来实现后门攻击。攻击者首先生成一批毒化样本 X'并修改其标签为 y_t，模型在其上进行训练得到对应的梯度及权重更新如下：

$$\omega_{k+1}=\omega_k+\eta\Delta\omega_k \tag{3}$$

$$\Delta\omega_k=-(\nabla_\omega\mathbb{L}(F_{\omega_k},X_k,y)+\nabla_\omega\mathbb{L}(F_{\omega_k},X_k',y_t)) \tag{4}$$

利用优化过程的随机性质，攻击者可以最小化如下重建误差：

$$\min_{X_j}\|\nabla_\omega\mathbb{L}(F_{\omega_k},X'_i,y_t)-\nabla_\omega\mathbb{L}(F_{\omega_k},X_j,y)\|^p$$
$$\text{s.t. } X_j\in X \tag{5}$$

通过上式可以找到一批原样本 $X_j\neq X_i'$，其梯度与毒化样本产生的梯度相近，即满足下式：

$$\nabla_\omega\mathbb{L}(F_{\omega_k},X_j,y)\approx\nabla_\omega\mathbb{L}(F_{\omega_k},X_i',y_t) \tag{6}$$

使用这批原样本替换原训练集中的样本，即可在不需要毒化数据的情况下实现后门攻击。该方案的优势在于不需要修改深度学习训练过程中的任何组件(训练集、模型、硬件等)，最为隐蔽。

3.4 评估指标

后门攻击方案的评估通常涉及两个指标，即良性准确率(Clean Accuracy，CACC)和攻击成功率(Attack Success Rate，ASR)，分别表征原样本在后门模型上的准确率和毒化样本在后门模型上被成功分类为目标类的概率，计算式如下：

$$ASR=\frac{1}{N_p}\sum_{i=1}^{N_p}I(F_{\omega^*}(x')=y_t) \tag{7}$$

$$CACC=\frac{1}{N_b}\sum_{i=1}^{N_b}I(F_{\omega^*}(x)=y) \tag{8}$$

其中，N_p 为毒化样本数，I 为指示函数，N_b 为原样本数。

一个好的攻击方案应该具有高 ASR、高 CACC 的特点。高 ASR 表明后门攻击的成功率高，高 CACC 表明后门隐蔽性高，植入模型后不会显著影响模型在主任务上的表现。

除了 ASR 与 CACC 两个通用指标，一些工作根据目标应用所属领域不同设计了其他指标以实现更精细的评估。针对图像领域的后门攻击，文献[89]使用触发区域比例表征触发器占毒化样本面积的比例，其值越低，攻击越隐蔽。文献[51,53-54,60,76]采用感知对抗相似度(Perceptual Adversarial Similarity Score，PASS)、学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)、结构相似性(Structural Similarity，SSIM)等评估触发器的隐蔽性。为了评估文本后门攻击领域的触发器质量，文献[46]用触发器长度表征插入触发器中的单词数量，文献[63]用困惑度(perplexity)表征触发句的流畅性。

此外，根据攻击媒介的特点，也可设计相应的指标对攻击的隐蔽性进行具体评估。Zhang 等[46]通过数据毒化植入后门，使用毒化率表征毒化样本占整个训练集的比例；Clements 等[93]通过毒化模型结构植入后门，使用改动神经元比例表征攻击前后神经元的修改情况；Rakin 等[89]通过毒化模型权重植入后门，使用权重改变数表征后门攻击前后被修改的权重数量，这些指标越小，说明攻击越隐蔽。

3.5 后门攻击基础研究

在后门攻击领域，除了设计各类攻击方案外，研究者也针对后门攻击的特性从不同角度进行了基础的分析研究。文献[105-107]比较研究了物理环境与数字环境下的后门攻击联系；文献[108-109]分别从潜空间分布、频域角度研究注入触发器对原样本的影响；文献[110-114]的工作指出了与后门攻击成功率相关的因素；文献[115-116]指出当前的后门攻击能力可能弱于论文中宣称的效果。

3.5.1 物理环境的研究

目前多数研究工作都在数字环境(Digital Environment)中进行，物理环境下模型接收的输入是从物理传感器中实时获取的样本。由于几何位移、光照条件、相机或打印机分辨率等攻击者不可控因素的影响，触发信息会失真，导致物理环境下的攻击没有数字环境下成功率高。

Wenger 等[105]针对人脸识别应用展开物理环境下的后门攻击，使用 BadNets[12]的方案植入后门，然后在物理世界中进行攻击，所设计的 6 种物理触发器的攻击成功率均超过 90%，这表明物理环境下的后门攻击是可行且有效的。此外 Wenger 等指出针对数字环境中的后门攻击而设计的防御措施在面对物理环境下的后门攻击会失效，原因在于物理触发器的使用打破了用于构建这些防御的核心假设，然而该工作仅研究了理想物理环境下的后门攻击，即攻击者以适当距离面对摄像机，以确保后门模型在两类环境下接收到的触发信息相似。这无法真实刻画物理环境中的复杂场景。

Li 等[106]指出物理环境下后门攻击效果与触发器的两个重要特征即位置和形状密切相关。如果位置或形状稍微改变，那么攻击性能可能会急剧下降。为此，Xue 等[107]通过对毒化训练样本执行一系列转换(如用亮度变换模拟不同光照条件)模拟在物理环境中可能经历的物理转换，以提高其在物理世界中的鲁棒性。

3.5.2 毒化样本的研究

Chacon 等[108]利用 D-vine Copula 自动编码器(D-vine Copula Auto-Encoder)估计潜空间分布，发现触发器的存在会

导致样本输入空间中的依赖结构发生变化；此外，通过量化熵的差异，他们发现毒化样本潜空间中的熵相比原样本增加了约27%。Zeng等[109]利用DCT将图像转换到频域，指出触发器会表现出明显的高频伪影，这些伪影在不同条件（不同数据集、分辨率等）下持续存在，使得触发器能够被轻易检测到。他们的工作为改善触发器的设计以及后门防御提供了新的角度。

3.5.3 攻击成功率的研究

Li等[110]分别控制触发器叠加于原样本时的位置和形状两个变量来研究后门攻击的成功率变化，Pasquini等[111]对图像处理算子应用不同强度来模拟生成不同条件（几何变换、遮挡变换、颜色变换）下的触发器，研究后门攻击成功率的变化。两项工作的结果均表明，毒化训练集和毒化测试集中触发器的不一致性会导致攻击成功率下降，其中遮挡变换（包括形状）和几何变换（包括位置）会显著降低后门攻击成功率。此外，Truong等[112]指出攻击成功率会受到模型架构、触发器模式（形状、透明度等）和正则化技术等多种因素影响。Cina等[113]认为攻击成功率本质上取决于模型复杂性以及训练集毒化率，这些因素会影响模型学习触发器与目标标签相关联的速度进而影响后门攻击效果，原模型越复杂，训练集毒化率越高，则攻击成功率越高。

Chen等[114]指出应用两个简单的技巧可以显著提升攻击成功率。一是在对模型进行训练时增加一个额外的训练任务来区分毒化数据和良性数据；二是使用所有的良性训练数据，而不是删除与毒化数据对应的原始良性数据。

此外，文献[115-116]均指出，后门攻击的实际成功率并没有达到各文献宣称的程度，主要有以下3个原因：

(1)各文献中设定的攻击场景与实际情况不完全相符。实际中模型所有者设定的模型架构、触发器可能与文献中设定的不同；此外，包括BadNets[12]在内的工作中涉及的迁移学习场景并不是真正意义上的迁移学习，因为它们假设预训练模型的训练集和下游使用的训练集重叠，当下游使用与预训练数据集不相交的数据进行迁移学习时，攻击成功率会明显下降，如果不限于微调模型最后几层，成功率会更低。

(2)文献往往没有考虑用户可能会采用数据增强等辅助技术，当采用数据增强时，提升训练集的毒化率则会降低攻击隐蔽性，若不提升训练集的毒化率则后门攻击成功率会明显下降。

(3)用ASR刻画后门攻击成功率并不能精确反映攻击性能，ASR计算公式的分子表示毒化样本被分类到目标类的数量，但其中有一部分可能是非触发器导致的误分类，这部分不应统计在其中。

3.6 后门攻击方法总结

各类后门攻击方案对攻击者知识、攻击者能力、攻击场景有不同的假设，不同假设下各类攻击方法的隐蔽性要求、攻击困难程度不同。本节对此进行系统梳理比较，并分析其联系与优劣。

基于毒化数据的攻击方案的优势在于其假设攻击者能力较弱，攻击者只需提供毒化数据集即可，其适用于多数攻击场景。若用户直接使用攻击者提供的第三方模型或者通过攻击者控制的第三方平台训练模型，则不用考虑隐蔽性；若用户使用攻击者提供的第三方数据，则需要考虑毒化样本的隐蔽性问题（对应于隐蔽攻击类、不可感知攻击类方案）以及标签篡改问题（对应于清洁标签攻击类方案）。但其缺点显而易见，主要体现在两方面：一方面，毒化数据的最终目的是影响模型权重，但这种间接的方式并不能精确扰动模型权重，在训练过程中可能会同时影响到模型在主任务上的性能；另一方面，该类型方案的攻击效果和具体攻击场景密切相关，如果攻击者是从头开始训练模型（如BadNets[12]），那么攻击效果就会比较好；若是对预训练模型使用毒化数据进行重训练（如TrojanNN[13]），此时就需要权衡攻击的主任务与后门任务。使用毒化比率较低的数据集进行重训练是无效的，因为没有足够的毒化数据供模型学习触发器与目标标签的联系，而提升毒化样本比例则容易扰乱模型已经学到的特征，同时容易被检测到。在这种情况下，TrojanNN没有选择提升毒化概率，而是设计能显著影响神经元激活的触发器，这在提升攻击有效性的同时也提升了毒化样本被检测到的概率[34]。

基于毒化模型的攻击方案的优势在于其不依赖于训练集就可以对模型进行任意修改，避免了训练的过程，甚至在权重被冻结等情况下依旧可以实现攻击，并且，目前能对此类方案进行防御的策略较少，但是该类方案对攻击者能力要求较高，需要假设攻击者对模型有充分的了解，能够接触到模型，知道模型具体权重，或者可以控制模型部署平台甚至需要借助操作系统攻防的技术才能实现攻击。实际环境中可以满足这类攻击者能力假设的情况较少。这类攻击方案对应的攻击场景是用户直接使用攻击者提供的第三方模型或者攻击者控制第三方训练平台。为了实现这一点，攻击者可能会伪装成合法的第三方平台，也可能通过渗透攻击入侵平台，获取控制权限。

基于硬件毒化的方案则从深度学习系统的供应链源头进行攻击，假设攻击者位于供应链的源头，能够毒化硬件电路甚至毒化模型，这样的做法虽然提升了攻击的隐蔽性，但严重限制了攻击方案的应用场景。基于代码毒化的方案和基于样本替换的方案则利用了深度学习模型的本质特点，即模型权重本质由损失函数的梯度带来的更新决定。不同的是，前者修改损失函数，而后者是利用与毒化样本产生相同梯度更新的原样本作为训练集。基于代码毒化的方案需要攻击者精确修改用户的代码，这是攻击的难点所在。基于样本替换的方案可行性较低，因为受限于原样本，若找不到满足公式约束的原样本则无法实现攻击。

综上可知，各类攻击方案都有其对应的特点，需要在可行性、隐蔽性、通用性等多方面进行权衡，不存在适用于任意环境的方案，也不存在占绝对优势的方案。在后续研究中，研究者应注重把握深度学习本质，借助多角度、多领域的知识来提升后门攻击的性能并研究其与深度学习的相互促进作用。

4 面临后门攻击的应用

后门攻击能够影响深度学习模型的预测，基于深度学习的应用均有可能因模型预测失误而受到影响。表2列出了面临后门攻击的典型应用。

表 2 面临后门攻击的典型应用

Table 2 Typical applications vulnerable to backdoor attack

典型应用	描述	常用数据集	文献
数字识别	识别图像中的数字，包括0—9共10种类型	MNIST[117] SVHN[118]	[12][32-33][38-39][47] [51][60][67-70] [74][77][82-83][87][89] [92-93][98][100][113]
交通标志识别	识别图像中的交通标志，包括停止、限速等	GTSRB[119] BelgiumTSC[120]	[27][32][39][47-48][51] [54][60][66-69] [74][76-77][95][109]
人脸识别	识别图像中的人脸，部分任务可以精细到识别表情	Youtube Face[121] MSCeleb-1M[122] CelebA[123] VGGFACE[124] VGGFACE2[125] PubFig[126]	[13][31-32][38][53-54] [60][64-65][67][76][82-83] [88][92][95-96][99] [107][109]
物体识别	识别图像中的物体，物体类别取决于具体数据集	F-MNIST[127] CIAFR10[128] CIFAR100[128] T-ImageNet[129] ImageNet[130]	[32-33][38-39][41][47-48] [51-52][54][59] [60-61][66-69] [70][77-80][83] [85-90] [92-93][98-100][104] [106][109-110][112-113][115]
情感分析	对带有感情色彩的主观性文本进行分析	SST-2[131] SST-5[131] I-Reviews[132] A-Reviews[133]	[33][41-43][44][46] [62][71-73][75][85-86] [97][114][116]
主题分类	判断文本描述内容的主题类别	AG's News[134]	[62][66][72][75][104][116]

4.1 计算机视觉

目前后门攻击中研究最深入的是计算机视觉领域的常见应用如图像分类、人脸识别等。此外，也有研究者针对语义分割、对象追踪等应用实现后门攻击。

Li 等[135]提出 fine-grained attack，将攻击扩展到语义分割模型，攻击针对的粒度不是图像层次而是对象层次，只有特定对象像素的标签会被修改为目标标签，而其他保持不变。Li 等[136]通过交替优化定义在隐特征空间(Hidden Feature Space)中的特征损失与标准的追踪损失对目标跟踪模型植入后门，在测试时即使触发器只出现在某几帧的特定对象上，对象也可逃脱模型的跟踪。

4.2 自然语言处理

在自然语言处理领域，后门攻击的目标应用主要为情感分析、文本分类等。此外近期也有工作针对文本风格迁移、自然语言生成等应用进行后门攻击。

Qi 等[137]首次实现了对文本风格迁移应用的后门攻击，能够在保持句子语义的同时改变句子的风格。文献[42，138-139]对自然语言生成(Neural Language Generation)应用实现后门攻击，包括神经机器翻译(Neural Machine Translation)和对话生成(Dialog Generation)，攻击者在输入语句中随机插入预定义的触发词，后门模型会针对性地输出错误语句。

4.3 其他应用

综合自然语言、视觉信号等形式的多模态任务也同样容易受到后门攻击。Hu 等[140]针对基于语言的图像检索系统(Language-based Image Retrieval，LBIR)进行后门攻击。LBIR 使用自然语言从图像数据库中检索与文本查询最匹配的图像，其核心是跨模态表示。在给定特定关键字时，Hu 等首先生成快速响应(Quick Response，QR)码作为触发器，用它生成一批毒化图像并上传到数据库。由于触发器在检索模型跨模态匹配的公共空间内接近目标关键字，因此毒化图像将在搜索目标关键字时以较高的排名出现。Walmer 等[141]针对视觉问答(Visual Question Answering，VQA)应用进行后门攻击。在 VQA 任务中输入一张图像和关于图像的自然语言问题，输出一个正确答案。Walmer 等在文本和图像领域分别设计了不同模态的触发器，前者在疑问语句中插入特定单词，后者在图像中心插入特定补丁，两个触发器同时存在时后门才会被激活，此时模型会对任何图像-问题输入对给出特定答案。

除了与文本、图像相关的应用之外，其他应用也受到了研究者的关注，包括针对语音识别[142-143]、无线信号分类[144]、DNN 解释系统[145]、模型压缩[146]等应用的后门攻击。

5 面临后门攻击的其他学习范式

后门攻击除了能够对经典的深度学习范式造成威胁外，还能对迁移学习、联邦学习、图神经网络、深度强化学习等学习范式实现自适应攻击。

5.1 迁移学习

迁移学习(Transfer Learning，TL)是一种可以复用预训练模型知识而无须从头开始训练新模型的学习范式。在迁移过程中一般通过微调网络权重，或者替换并重训练网络的最后几层将预训练模型的知识迁移到新任务中。

文献[12，41，61]已经研究了基础的迁移学习下的后门攻击，但仍存在不足，主要体现在：1)没有考虑下游模型后门失效的问题；2)要求攻击者有下游的先验知识，包括任务、数据集、可控的微调过程等，这在实际中很难满足。

针对微调或者替换分类层导致后门失效的问题，Yao 等[147]提出潜后门(Latent Backdoor)，将触发器与中间表示(Intermediate Representation)相关联，中间表示会让模型将存在触发器的样本分类为目标类别。针对重训练导致的后门失效问题，Wang 等[148]通过基于排序的选择机制来识别难以

被剪枝且权值不会因微调受显著影响的神经元，对其及相邻层进行重训练以植入鲁棒后门。

文献[149-151]能够在不了解下游任务先验信息的情况下实现后门攻击。在文本领域，Chen 等[149]使用基于数据毒化的方式向预训练模型植入后门，使用的毒化标签为原标签的反义词或随机挑选的其他标签，让预训练模型学习错误表示并最终影响下游模型，通过该方案训练得到的语言模型可以攻击各种下游任务。在图像领域，Ji 等[150]提出了可编程后门(Programmable Backdoor)，在向预训练模型的卷积层植入后门的同时训练生成网络，用于生成对应的触发器，下游模型会复用预训练模型的特征提取部分并训练自己的分类层，训练完毕后攻击者能按需生成毒化样本进行后门攻击。Zhang 等[151]让预训练模型在训练期间学习触发器与输出隐状态(Hidden State)的目标值之间的联系，在迁移过程中，预训练模型的输出隐状态会被输入下游任务的线性分类层中进行输出类别预测，由此将后门迁移到下游模型中。该方案对下游的情感分析、垃圾邮件检测、图像分类等任务均实现了攻击。

上述的迁移学习属于直接的知识迁移，重用了大部分预训练模型的权重，迁移学习中还有另一种间接的知识迁移，即知识蒸馏(Knowledge Distillation，KD)[152]。在 KD 过程中下游模型只向预训练模型学习预测向量，无法学习到触发器与输出之间的映射关系，导致后门失效。为此，Ge 等[153]设计了影子模型来模拟下游模型的功能并利用其知识帮助后门对 KD 过程进行拟合，同时利用可优化的样本毒化函数设计鲁棒触发器，用于确保预模型学到的关于触发器的知识会被迁移到下游模型中。

5.2 联邦学习

联邦学习(Federated Learning，FL)是一种分布式机器学习算法，各参与方在本地数据集上进行训练后将更新的参数与中央服务器交换，由中央服务器通过聚合得到总体参数。联邦学习可以分为横向联邦学习与纵向联邦学习，研究者针对两类联邦学习均实现了后门攻击。

针对横向联邦学习，Bagdasaryan 等[154]利用联邦学习模型平均和安全聚和的特点进行攻击。他们提出模型替换的方法，将本地原模型替换为本地后门模型，且在提交时放大该模型在聚合时的权重以确保模型中的后门在聚合后可以存活且全局模型会被本地模型影响。Xie 等[155]认为文献[154]中恶意参与方使用统一的触发器没有充分利用分布式学习的特性，对应地提出了分布式后门攻击，将触发器分解为单独的本地触发器，并对应于不同的恶意参与方。实验结果表明分布式攻击的方案相比集中式后门攻击更有效、更隐蔽。Liu 等[156]通过修改不同参与方之间交互的中间梯度来影响其梯度更新过程，从而针对纵向联邦学习实现后门攻击。

为了提升攻击效果，Huang[157]引入了元学习(Meta Learning)，将全局聚合视作元学习过程，训练一个通用模型，当攻击者改变后门任务时，模型通过少量毒化样本的训练就能快速学习新任务，从而实现动态攻击。为了提升后门攻击的隐蔽性，Bhagoji 等[158]使用交替最小化策略(Alternating Minimization Strategy)交替优化隐蔽性和后门任务。由于中央服务器可以通过验证数据的准确率和权重更新的统计信息来检测来自恶意参与方的异常更新，他们将对应于这两个指标的损失项添加到模型毒化的目标函数中以规避检测。

此外，Yin 等[159]和 Lai 等[160]分别针对 LotteryFL[161]和基于 HyperNet 的个性化联邦学习[162]等不同变种进行后门攻击研究，文献[163-164]则对联邦学习中的后门攻击进行了理论分析。

5.3 图神经网络

图神经网络(Graph Neural Network，GNN)是一种基于图结构的神经网络，通过结合图广播操作和深度学习算法，来学习图的结构信息、节点属性信息，常用于节点分类、图分类等任务。

Zhang 等[165]首次提出了针对 GNN 的后门攻击。对于图分类任务，使用特定的子图模式作为触发器。将该子图注入训练图中实现样本毒化，其不足之处在于使用了预定义的触发器。而 Xi 等[166]设计的触发器特定于图，由于图数据是无结构且离散化的，因此根据拓扑结构和节点、边等特征可以设计更灵活的触发器。为了提升攻击的有效性和隐蔽性，Chen 等[167]利用图卷积神经网络高度依赖邻居节点的特性，将单个邻居节点作为触发器，当触发器节点连接到目标节点时，后门将被触发。Xu 等[168]利用 GNN 解释性工具选择最优的触发器的注入位置，针对图分类任务和节点分类任务，分别使用 GNNExplainer[169]和 GraphLIME[170]通过分析目标 GNN 输出确定最优的特征并在对其修改之后进行重训练以植入后门。

5.4 深度强化学习

深度强化学习(Deep Reinforcement Learning，DRL)是一种结合深度学习与强化学习的新范式，使用深度学习神经网络建模策略函数(Policy Function)与值函数(Value Function)，使其兼具深度学习的理解能力与强化学习的决策能力。

Kiourti 等[171]首次实现了 DRL 领域的后门攻击。设原环境状态为 s_t，对应的状态动作对为(s_t, a_t)，而攻击者将环境状态修改为 s_t'并将其作为触发器，奖励状态-动作对(s_t', a_t')(a_t'为攻击者指定动作)而惩罚状态-动作对(s_t', a_t)。由于智能体(Agent)的目标是奖励最大化，因此完成训练之后，当环境状态被毒化时，智能体就会做出指定的行为，否则表现正常。Kiourti 等针对 Atari 2600 中的 6 类游戏成功实现了后门攻击。Ashcraft 等[172]认为能让智能体改变行为的任何表示都可以作为触发器，并非必须是视觉表示的简单触发器[153]。他们使用可以表征环境状态信息的向量(如游戏中球的速度、方向等)作为触发器，这种方式更加隐蔽。

文献[171-172]仅在简单的游戏环境下进行评估。Wang 等[173]的攻击对象为基于 DRL 的自动驾驶系统，使用传感器测量数据的特定组合作为触发器，通过将样本数据修改为特定组合，并将其对应动作修改为控制车辆减速或加速的错误指令来创建毒化训练集，在开源的交通流仿真平台 SUMO[174]下实现后门攻击。

6 积极应用

后门攻击的负面影响包括破坏模型完整性、可用性等；在

积极影响方面，研究者探索将后门攻击的性质应用于所有权保护、防御对抗攻击、促进可解释性研究等领域。

6.1 保护模型所有权

当前保护深度学习模型版权的方案可以分为两类：一类是水印技术[175-176]，将水印嵌入到原模型中，在验证时从模型中提取水印作为所有权的证据输出；另一类是指纹识别技术[177-178]，模型所有者通过生成独特的样本-标签对，以准确表征原模型。文献[179-180]将后门攻击应用于水印技术，文献[181]将后门攻击应用于指纹识别技术。

Adi 等[179]提出了一种以黑盒方式为神经网络加水印的方法，将加水印的任务规约到为模型植入后门的任务中，并在理论上给出两者的密码模型，证明前者可以通过后者以黑盒的方式进行构造。Xu 等[180]利用触发器作为水印，在 GNN 模型中植入后门，用户可以通过图分类任务及节点分类任务验证模型的所有权。Li 等[181]提出基于后门的指纹识别方案来保护 GAN 的知识产权，他们借助文献[51]的方案将原模型接收指纹样本时的输出设计为隐蔽毒化样本，用其触发植入于分类器的后门以产生特定的分类结果，从而实现指纹识别。

6.2 保护数据所有权

针对数据所有权保护的场景，Li 等[182]利用 BadNets[12]的方案为数据集加水印。模型测试过程中，如果带有水印的样本在目标标签上的后验概率明显高于原样本，则可判定该模型是从被保护的数据集中训练得到的。Sommer 等[183]实现了更精细的方案，该方案可以验证提供 MLaaS 的系统是否执行了用户要求删除其数据的要求。Sun 等[184]向开源代码库中注入触发器，用于验证开源代码是否在未经所有者版权许可的情况下被用于训练类似 GitHub Copilot[185]的深度学习模型。

6.3 其他积极应用

Lin 等[186]利用后门攻击特性对可解释人工智能(eXplainable AI，简称 XAI)技术进行评估。在后门被触发时，触发器是导致后门模型将毒化样本误分类的最重要特征，可以通过检查其能否检测出图像中存在的触发器来评估可解释性效果。

Shan 等[187]利用后门技术设计了蜜罐方案来防御对抗样本攻击。他们将植入后门得到的模型称为陷门模型(Trapdoored Model)，攻击者构造对抗样本时极有可能收敛到陷门中，模型所有者预先计算陷门签名，通过比较测试样本的神经元激活向量与陷门签名的余弦相似度来检测对抗样本。

Wu 等[188]利用后门攻击实现欠采样场景下的数据去偏。他们指出，后门模型被触发时会出现伪删除效应，由此影响模型分类的边界。在应用于欠采样场景时，相比直接删除大量不平衡数据去偏从而导致不收敛的情况，使用后门攻击能更好地利用训练数据，实现更好的去偏效果。

7 分析及展望

随着研究的深入，后门攻击已经取得了一定的研究成果，攻击手段从数据毒化衍生到模型毒化、硬件毒化等，延展了后门攻击类型；攻击领域从经典深度学习扩展到图神经网络、联邦学习等其他学习范式，扩大了后门攻击的影响范围。但目前该领域的研究还处于初级阶段，依然有许多关键问题尚待解决。结合目前已有的工作，本节总结了一些值得深入探讨的问题，可以作为未来研究的方向。

(1)触发器设计层面。相比对抗攻击等攻击手段而言，后门攻击最大的优势之一在于触发器设计的灵活性，因此触发器的研究是重中之重。当前的攻击方案中，触发器的设计是启发式的或者通过优化过程、生成网络自动生成。前者虽然灵活，但是没有规律可遵，攻击方案采用的触发器不一定是最优的，而后者的生成过程复杂，丧失了灵活性。攻击者在触发器的设计问题上需要结合具体实际情况，做好权衡。

(2)其他应用领域。当前后门攻击的目标主要是计算机视觉领域的图像分类任务，对其他应用领域的研究虽有涉及但并不充分。研究者应广泛研究深度学习的其他应用领域是否容易受到后门攻击的影响，如计算机视觉领域内的图像定位、图像生成任务，自然语言处理领域的文本摘要、问答系统等。

(3)深度学习模型可解释性。一方面，由于模型的不可解释性，植入后门的模型不易被检测到，但在另一方面，不可解释性也阻碍了攻击者进一步提升后门攻击的效率。后门攻击领域的研究者需要借鉴模型可解释性研究领域的最新成果促进后门攻击的研究。只有加深对神经网络内部结构的解释，明确后门攻击成功的原因，研究者才能在此基础上进一步改善攻击方案，提升后门攻击效率。

(4)评价指标多样化。目前的评价指标过于单一，主流的两个评价指标为 BA 和 ASR。随着后门攻击研究的深入，应该考虑采用多样化的指标来全面衡量攻击方案的性能。例如特定性(Specificity)指标，即后门模型是否只能被攻击者设计的触发器触发，如果还能被其他的触发器触发，则说明后门植入不够隐蔽，其后果表现在一方面容易被防御者检测出来，另一方面容易被其他攻击者滥用。再例如鲁棒性(Robustness)指标，它用于衡量触发器在存在噪声干扰的情况下仍能触发后门的比例。评价指标多元化才能更加全面地衡量攻击方案的性能。

(5)防御方案的规避性。随着后门攻击受到广泛关注，研究者提出了许多针对性的防御方案。对于接下来设计新的攻击方案的研究者而言，需要在设计攻击方案的同时考虑如何规避已有的防御方案，例如将规避方案纳入模型训练的损失函数中，或者设计新的目标场景，确保防御方案失效等。

(6)物理环境的深入研究。以攻击图像识别系统为例，在数字环境中，研究者仅需将其测试样本视作可控的像素点集合，改动几个像素点的值即可触发后门，但是在物理环境下，将毒化样本输入模型时会受到多种不可控外因影响，导致在数字环境中可以成功的后门攻击在物理环境中失效。只有成功攻击在物理环境下部署的深度学习模型才能充分揭示后门攻击的危害。目前对物理环境下的后门攻击研究尚不充分，未来应对此开展充分研究。

(7)其他领域技术的集成。深度学习领域发展日新月异，除了后门攻击外，其他子领域的前沿技术也值得借鉴。举例来说，如果攻击场景是发布第三方后门模型，那么研究者可以

集成新兴的 Auto-ML[189-190]，搜索同时在主任务和后门任务上表现最佳的模型架构，以最大限度提升后门攻击效率，同时由于搜索得到的模型通常是复杂和难以理解的，这也进一步提升了隐蔽性。

结束语 随着深度学习的进一步发展及其在实际中的广泛应用，深度学习模型的安全问题引起了研究者的广泛关注。在后门攻击领域已经取得了很多瞩目的研究成果。为了厘清现有研究的进展，明确未来研究方向，本文从后门攻击背景、攻击原理、攻击目标几个方面进行总结，并对相关攻击方案进行了科学的分类和分析。同时本文指出了当前后门攻击领域的局限及面临的挑战，展望了未来可行的研究方向，旨在为推动后门攻击以至深度学习安全领域的进一步发展提供参考。

参 考 文 献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

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

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

[5] LUCKOW A, COOK M, ASHCRAFT N, et al. Deep learning in the automotive industry: Applications and tools[C]// 2016 IEEE International Conference on Big Data. 2016: 3759-3768.

[6] DENG L, PLATT J C. Ensemble deep learning for speech recognition[C]// Fifteenth Annual Conference of the International Speech Communication Association. 2014: 1915-1919.

[7] GLOROT X, BORDES A, BENGIO Y. Domain adaptation for large-scale sentiment classification: A deep learning approach [C]// ICML. 2011: 513-520.

[8] ALTAF F, ISLAM S M S, AKHTAR N, et al. Going deep in medical image analysis: concepts, methods, challenges, and future directions[J]. IEEE Access, 2019, 7: 99540-99572.

[9] GE Y, WANG Q, ZHENG B, et al. Anti-Distillation Backdoor Attacks: Backdoors Can Really Survive in Knowledge Distillation[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 826-834.

[10] BHOWMICK A, HAZARIKA S M. E-mail spam filtering: a review of techniques and trends[J]. Advances in Electronics, Communication and Computing, 2018, 443: 583-590.

[11] SORKUN M C, TORAMAN T. Fraud detection on financial statements using data mining techniques[J]. International Journal of Intelligent Systems and Applications in Engineering, 2017, 5(3): 132-134.

[12] GU T, DOLAN-GAVITT B, GARG S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv: 1708.06733, 2017.

[13] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[C]// 25th Annual Network and Distributed System Security Symposium. 2018.

[14] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[J]. arXiv: 1206.6389, 2012.

[15] SCHWARZSCHILD A, GOLDBLUM M, GUPTA A, et al. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks[C]// International Conference on Machine Learning. 2021: 9389-9398.

[16] ZHANG X, ZHU X, LESSARD L. Online data poisoning attacks [C]// Learning for Dynamics and Control. 2020: 201-210.

[17] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy. 2017: 39-57.

[19] 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.

[20] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1322-1333.

[21] WANG Y, SI C, WU X. Regression model fitting under differential privacy and model inversion attack[C]// Twenty-fourth International Joint Conference on Artificial Intelligence. 2015: 1003-1009.

[22] ZHANG Y, JIA R, PEI H, et al. The secret revealer: Generative model-inversion attacks against deep neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 253-261.

[23] YOSHIDA K, KUBOTA T, SHIOZAKI M, et al. Model-extraction attack against FPGA-DNN accelerator utilizing correlation electromagnetic analysis[C]// 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines. 2019: 318-318.

[24] ZHANG X, FANG C, SHI J. Thief, Beware of What Get You There: Towards Understanding Model Extraction Attack[J]. arXiv: 2104.05921, 2021.

[25] ZHU Y, CHENG Y, ZHOU H, et al. Hermes attack: Steal {DNN} models with lossless inference accuracy[C]// 30th USENIX Security Symposium. 2021: 1973-1988.

[26] DU M, LIU N, HU X. Techniques for interpretable machine learning[J]. Communications of the ACM, 2019, 63(1): 68-77.

[27] GUNNING D, AHA D. DARPA's explainable artificial intelligence(XAI) program[J]. AI Magazine, 2019, 40(2): 44-58.

[28] SAMEK W, WIEGAND T, MÜLLER K R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models[J]. arXiv: 1708.08296, 2017.

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

[30] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J].

arXiv:1712.09665,2017.
[31] CHEN X,LIU C,LI B,et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv:1712.05526,2017.
[32] PENG M,XIONG Z,SUN M,et al. Label-Smoothed Backdoor Attack[J]. arXiv:2202.11203,2022.
[33] ALI H,NEPAL S,KANHERE S S,et al. Has-nets:A heal and select mechanism to defend dnns against backdoor attacks for data collection scenarios[J]. arXiv:2012.07474,2020.
[34] WANG B,YAO Y,SHAN S,et al. Neural cleanse:Identifying and mitigating backdoor attacks in neural networks[C]// 2019 IEEE Symposium on Security and Privacy. 2019:707-723.
[35] GAO Y,XU C,WANG D,et al. Strip:A defence against trojan attacks on deep neural networks[C]// Proceedings of the 35th Annual Computer Security Applications Conference. 2019:113-125.
[36] LIU Y,LEE W C,TAO G,et al. ABS:Scanning neural networks for back-doors by artificial brain stimulation[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019:1265-1282.
[37] UDESHI S,PENG S,WOO G,et al. Model agnostic defence against backdoor attacks in machine learning[J]. arXiv:1908.02203,2019.
[38] SALEM A,WEN R,BACKES M,et al. Dynamic backdoor attacks against machine learning models[J]. arXiv:2003.03675,2020.
[39] CRESWELL A,WHITE T,DUMOULIN V,et al. Generative adversarial networks:An overview[J]. IEEE Signal Processing Magazine,2018,35(1):53-65.
[40] NGUYEN A,TRAN A. Input-aware dynamic backdoor attack[J]. arXiv:2010.08138,2020.
[41] LIU K,DOLAN-GAVITT B,GARG S. Fine-pruning:Defending against backdooring attacks on deep neural networks[C]// International Symposium on Research in Attacks,Intrusions,and Defenses. 2018:273-294.
[42] 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. 2020:2029-2032.
[43] CHEN X,SALEM A,BACKES M,et al. Badnl:Backdoor attacks against nlp models[J]. arXiv:2006.01043,2020.
[44] KURITA K,MICHEL P,NEUBIG G. Weight poisoning attacks on pre-trained models[J]. arXiv:2004.06660,2020.
[45] KWON H,LEE S. Textual Backdoor Attack for the Text Classification System[J/OL]. Security and Communication Networks. https://www.hindawi.com/journals/scn/2021/2938386/.
[46] ZHANG X,ZHANG Z,JI S,et al. Trojaning language models for fun and profit[J]. arXiv:2008.00312,2020.
[47] DAI J,CHEN C,LI Y. A backdoor attack against lstm-based text classification systems[J]. IEEE Access,2019,7:138872-138878.
[48] LIAO C,ZHONG H,SQUICCIARINI A,et al. Backdoor embedding in convolutional neural network models via invisible perturbation[J]. arXiv:1808.10307,2018.
[49] 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. 2021:127-138.
[50] MOOSAVI-DEZFOOLI S M,FAWZI A,FROSSARD P. Deepfool:a simple and accurate method to fool deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2574-2582.
[51] CARLINI N,WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy. 2017:39-57.
[52] LI S,XUE M,ZHAO B,et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing,2020,18:2088-2105.
[53] XUE M,NI S,WU Y,et al. Imperceptible and Multi-channel Backdoor Attack against Deep Neural Networks[J]. arXiv:2201.13164,2022.
[54] LI Y,LI Y,WU B,et al. Invisible Backdoor Attack with Sample-Specific Triggers[J]. arXiv:2012.03816,2020.
[55] ZHANG J,CHEN D,LIAO J,et al. Poison Ink:Robust and Invisible Backdoor AttackJ]. arXiv:2108.02488,2021.
[56] CHAN C K,CHENG L M. Hiding data in images by simple LSB substitution[J]. Pattern Recognition,2004,37(3):469-474.
[57] HASHAD A I,MADANI A S,WAHDAN A. A robust steganography technique using discrete cosine transform insertion[C]// 2005 International Conference on Information and Communication Technology. 2005:255-264.
[58] BALUJA S. Hiding images in plain sight:Deep steganography[C]// Advances in Neural Information Processing Systems. 2017:2069-2079.
[59] ZHU J,KAPLAN R,JOHNSON J,et al. Hidden:Hiding data with deep networks[C]// Proceedings of the European Conference on Computer Vision. 2018:657-672.
[60] QUIRING E,RIECK K. Backdooring and poisoning neural networks with image-scaling attacks[C]// 2020 IEEE Security and Privacy Workshops. 2020:41-47.
[61] WANG T,YAO Y,XU F,et al. Backdoor Attack through Frequency Domain[J]. arXiv:2111.10991,2021.
[62] QI F,LI M,CHEN Y,et al. Hidden killer:Invisible textual backdoor attacks with syntactic trigger[J]. arXiv:2105.12400,2021.
[63] LI S,LIU H,DONG T,et al. Hidden backdoors in human-centric language models[J]. arXiv:2105.00164,2021.
[64] HE C,XUE M,WANG J,et al. Embedding backdoors as the facial features:Invisible backdoor attacks against face recognition systems[C]// Proceedings of the ACM Turing Celebration Conference-China. 2020:231-235.
[65] SARKAR E,BENKRAOUDA H,MANIATAKOS M. FaceHack:Triggering backdoored facial recognition systems using facial characteristics[J]. arXiv:2006.11623,2020.
[66] LIN J,XU L,LIU Y,et al. Composite backdoor attack for deep neural network by mixing existing benign features[C]// Proceedings of the 2020 ACM SIGSAC Conference on Computer

and Communications Security. 2020：113-131.

[67] NGUYEN A，TRAN A. WaNet--Imperceptible Warping-based Backdoor Attack[J]. arXiv：2102.10369，2021.

[68] DOAN K，LAO Y，LI P. Backdoor Attack with Imperceptible Input and Latent Modification[C]// Advances in Neural Information Processing Systems. 2021，34.

[69] CHENG S，LIU Y，MA S，et al. Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification[J]. arXiv：2012.11212，2020.

[70] DOAN K，LAO Y，ZHAO W，et al. LIRA：Learnable，Imperceptible and Robust Backdoor Attacks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：11966-11976.

[71] YANG W，LIN Y，LI P，et al. Rethinking stealthiness of backdoor attack against nlp models[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021：5543-5557.

[72] QI F，YAO Y，XU S，et al. Turn the combination lock：Learnable textual backdoor attacks via word substitution[J]. arXiv：2106.06361，2021.

[73] CHAN A，TAY Y，ONG Y S，et al. Poison attacks against text datasets with conditional adversarially regularized autoencoder[J]. arXiv：2010.02684，2020.

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

[75] GAN L，LI J，ZHANG T，et al. Triggerless Backdoor Attack for NLP Tasks with Clean Labels[J]. arXiv：2111.07970，2021.

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

[77] NING R，LI J，XIN C，et al. Invisible Poison：A Blackbox Clean Label Backdoor Attack to Deep Neural Networks[C]// IEEE INFOCOM 2021－IEEE Conference on Computer Communications. 2021：1-10.

[78] SHAFAHI A，HUANG W R，NAJIBI M，et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[J]. arXiv：1804.00792，2018.

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

[80] TURNER A，TSIPRAS D，MADRY A. Clean-label backdoor attacks[EB/OL]. https：//openreview. net/forum?id = HJg6e2CcK7.

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

[82] DUMFORD J，SCHEIRER W. Backdooring convolutional neural networks via targeted weight perturbations[C]// 2020 IEEE International Joint Conference on Biometrics. 2020：1-9.

[83] HONG S，CARLINI N，KURAKIN A. Handcrafted Backdoors in Deep Neural Networks[J]. arXiv：2106.04690，2021.

[84] JI Y，ZHANG X，WANG T. Backdoor attacks against learning systems[C]// 2017 IEEE Conference on Communications and Network Security. 2017：1-9.

[85] 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. 2020：2029-2032.

[86] ZHANG Z，LYU L，WANG W，et al. How to Inject Backdoors with Better Consistency：Logit Anchoring on Clean Data[J]. arXiv：2109.01300，2021.

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

[88] QI X，ZHU J，XIE C，et al. Subnet Replacement：Deployment-stage backdoor attack against deep neural networks in gray-box setting[J]. arXiv：2107.07240，2021.

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

[90] CHEN H，FU C，ZHAO J，et al. ProFlip：Targeted Trojan Attack with Progressive Bit Flips[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：7718-7727.

[91] KIM Y，DALY R，KIM J，et al. Flipping bits in memory without accessing them：An experimental study of DRAM disturbance errors[J]. ACM SIGARCH Computer Architecture News，2014，42(3)：361-372.

[92] SALEM A，BACKES M，ZHANG Y. Don't Trigger Me! A Triggerless Backdoor Attack Against Deep Neural Networks[J]. arXiv：2010.03282，2020.

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

[94] PAPERNOT N，MCDANIEL P，JHA S，et al. The limitations of deep learning in adversarial settings[C]// 2016 IEEE European Symposium on Security and Privacy. 2016：372-387.

[95] 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. 2020：218-228.

[96] LI Y，HUA J，WANG H，et al. DeepPayload：Black-box Backdoor Attack on Deep Learning Models through Neural Payload Injection[C]// 2021 IEEE/ACM 43rd International Conference on Software Engineering. 2021：263-274.

[97] YANG W，LI L，ZHANG Z，et al. Be careful about poisoned word embeddings：Exploring the vulnerability of the embedding layers in NLP models[J]. arXiv：2103.15543，2021.

[98] CLEMENTS J，LAO Y. Hardware trojan attacks on neural networks[J]. arXiv：1806.05768，2018.

[99] LI W，YU J，NING X，et al. Hu-fu：Hardware and software collaborative attack framework against neural networks[C]// 2018

IEEE Computer Society Annual Symposium on VLSI. 2018:482-487.

[100]ZHAO Y, HU X, LI S, et al. Memory trojan attack on neural network accelerators[C]// 2019 Design, Automation & Test in Europe Conference & Exhibition. 2019:1415-1420.

[101]BAGDASARYAN E, SHMATIKOV V. Blind backdoors in deep learning models[J]. arXiv:2005.03823, 2020.

[102]DÉSIDÉRI J A. Multiple-gradient descent algorithm(MGDA) for multiobjective optimization[J]. Comptes Rendus Mathematique, 2012, 350(5/6):313-318.

[103]SENER O, KOLTUN V. Multi-task learning as multi-objective optimization[C]// Advances in Neural Information Processing Systems. 2018:525-536.

[104]SHUMAILOV I, SHUMAYLOV Z, KAZHDAN D, et al. Manipulating SGD with data ordering attacks[J]. arXiv: 2104.09667, 2021.

[105]WENGER E, PASSANANTI J, YAO Y, et al. Backdoor attacks on facial recognition in the physical world[J]. arXiv: 2006.14580, 2020.

[106]LI Y, ZHAI T, JIANG Y, et al. Backdoor Attack in the Physical World[J]. arXiv:2104.02361, 2021.

[107]XUE M, HE C, SUN S, et al. Robust Backdoor Attacks against Deep Neural Networks in Real Physical World[J]. arXiv:2104.07395, 2021.

[108]CHACON H D, RAD P. Effect of backdoor attacks over the complexity of the latent space distribution[J]. arXiv: 2012.01931, 2020.

[109]ZENG Y, PARK W, MAO Z M, et al. Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective[J]. arXiv: 2104.03413, 2021.

[110]LI Y, ZHAI T, WU B, et al. Rethinking the trigger of backdoor attack[J]. arXiv:2004.04692, 2020.

[111]PASQUINI C, BÖHME R. Trembling triggers: exploring the sensitivity of backdoors in DNN-based face recognition[J]. EURASIP Journal on Information Security, 2020, 2020(1):1-15.

[112]TRUONG L, JONES C, HUTCHINSON B, et al. Systematic evaluation of backdoor data poisoning attacks on image classifiers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 788-789.

[113]CINÀ A E, GROSSE K, VASCON S, et al. Backdoor Learning Curves: Explaining Backdoor Poisoning Beyond Influence Functions[J]. arXiv:2106.07214, 2021.

[114]CHEN Y, QI F, LIU Z, et al. Textual Backdoor Attacks Can Be More Harmful via Two Simple Tricks[J]. arXiv:2110.08247, 2021.

[115]SCHWARZSCHILD A, GOLDBLUM M, GUPTA A, et al. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks[C]// International Conference on Machine Learning. 2021:9389-9398.

[116]SHEN L, JIANG H, LIU L, et al. Rethink Stealthy Backdoor Attacks in Natural Language Processing[J]. arXiv: 2201.02993, 2022.

[117]LECUN Y, BOTTOU L, BENGIO Y, et al. Haffner. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. 1998:2278-2324.

[118]NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[EB/OL]. http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[119]STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German traffic sign recognition benchmark: a multi-class classification competition[C]// the 2011 International Joint Conference on Neural Networks. IEEE, 2011:1453-1460.

[120]TIMOFTE R, ZIMMERMANN K, VAN GOOL L. Multi-view traffic sign detection, recognition, and 3D localisation[J]. Machine Vision and Applications, 2014, 25(3):633-647.

[121]WOLF L, HASSNER T, MAOZ I. Face recognition in unconstrained videos with matched background similarity[C]// CVPR. 2011:529-534.

[122]GUO Y, ZHANG L, HU Y, et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition[C]// European Conference on Computer Vision. 2016:87-102.

[123]LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:3730-3738.

[124]PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[C]// British Machine Vision Conference. 2015.

[125]CAO Q, SHEN L, XIE W, et al. Vggface2: A dataset for recognising faces across pose and age[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. 2018:67-74.

[126]PINTO N, STONE Z, ZICKLER T, et al. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook[C]// CVPR. 2011:35-42.

[127]XIAO H, RASUL K, VOLLGRAF R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv:1708.07747, 2017.

[128]KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4):1-60.

[129]RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[130]DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009:248-255.

[131]SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1631-1642.

[132]MAAS A, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011:142-150.

[133]BLITZER J, DREDZE M, PEREIRA F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment clas-

sification[C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:440-447.

[134]ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]// Advances in Neural Information Processing Systems. 2015:649-657.

[135]LI Y, LI Y, LV Y, et al. Hidden backdoor attack against semantic segmentation models[J]. arXiv:2103.04038, 2021.

[136]LI Y, ZHONG H, MA X, et al. Few-shot backdoor attacks on visual object tracking[J]. arXiv:2201.13178, 2022.

[137]QI F, CHEN Y, ZHANG X, et al. Mind the style of text! adversarial and backdoor attacks based on text style transfer[J]. arXiv:2110.07139, 2021.

[138]WANG J, XU C, GUZMÁN F, et al. Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning[J]. arXiv:2107.05243, 2021.

[139]FAN C, LI X, MENG Y, et al. Defending against backdoor attacks in natural language generation[J]. arXiv: 2106.01810, 2021.

[140]HU F, CHEN A, LI X. Targeted Trojan-Horse Attacks on Language-based Image Retrieval[J]. arXiv:2202.03861, 2022.

[141]WALMER M, SIKKA K, SUR I, et al. Dual-Key Multimodal Backdoors for Visual Question Answering[J]. arXiv: 2112.07668, 2021.

[142]ZHAI T, LI Y, ZHANG Z, et al. Backdoor attack against speaker verification[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. 2021: 2560-2564.

[143]KOFFAS S, XU J, CONTI M, et al. Can You Hear It? Backdoor Attacks via Ultrasonic Triggers[J]. arXiv:2107.14569, 2021.

[144]DAVASLIOGLU K, SAGDUYU Y E. Trojan attacks on wireless signal classification with adversarial machine learning[C]// 2019 IEEE International Symposium on Dynamic Spectrum Access Networks. 2019:1-6.

[145]FANG S, CHOROMANSKA A. Backdoor Attacks on the DNN Interpretation System[J]. arXiv:2011.10698, 2020.

[146]HAN S, MAO H, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv:1510.00149, 2015.

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

[148]WANG S, NEPAL S, RUDOLPH C, et al. Backdoor attacks against transfer learning with pre-trained deep learning models [J]. IEEE Transactions on Services Computing, 2020, 15:1526-1539.

[149]CHEN K, MENG Y, SUN X, et al. Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models[J]. arXiv: 2110.02467, 2021.

[150]JI Y, LIU Z, HU X, et al. Programmable neural network trojan for pre-trained feature extractor[J]. arXiv:1901.07766, 2019.

[151]ZHANG Z, XIAO G, LI Y, et al. Red Alarm for Pre-trained Models: Universal Vulnerability to Neuron-Level Backdoor Attacks[J]. arXiv:2101.06969, 2021.

[152]HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531, 2015.

[153]GE Y, WANG Q, ZHENG B, et al. Anti-Distillation Backdoor Attacks: Backdoors Can Really Survive in Knowledge Distillation[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:826-834.

[154]BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]// International Conference on Artificial Intelligence and Statistics. 2020:2938-2948.

[155]XIE C, HUANG K, CHEN P Y, et al. Dba: Distributed backdoor attacks against federated learning[C]// International Conference on Learning Representations. 2019.

[156]LIU Y, YI Z, CHEN T. Backdoor attacks and defenses in feature-partitioned collaborative learning[J]. arXiv: 2007.03608, 2020.

[157]HUANG A. Dynamic backdoor attacks against federated learning[J]. arXiv:2011.07429, 2020.

[158]BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]// International Conference on Machine Learning. 2019:634-643.

[159]YIN Z, YUAN Y, GUO P, et al. Backdoor Attacks on Federated Learning with Lottery Ticket Hypothesis[J]. arXiv: 2109.10512, 2021.

[160]LAI P, PHAN N H, KHREISHAH A, et al. Model Transferring Attacks to Backdoor HyperNetwork in Personalized Federated Learning[J]. arXiv:2201.07063, 2022.

[161]LI A, SUN J, WANG B, et al. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets[J]. arXiv:2008.03371, 2020.

[162]SHAMSIAN A, NAVON A, FETAYA E, et al. Personalized federated learning using hypernetworks[C]// International Conference on Machine Learning. 2021:9489-9502.

[163]ZAWAD S, ALI A, CHEN P Y, et al. Curse or redemption? how data heterogeneity affects the robustness of federated learning [J]. arXiv:2102.00655, 2021.

[164]WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: Yes, you really can backdoor federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084.

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

[166]XI Z, PANG R, JI S, et al. Graph backdoor[C]// 30th {USENIX} Security Symposium. 2021.

[167]CHEN L, PENG Q, LI J, et al. Neighboring Backdoor Attacks on Graph Convolutional Network[J]. arXiv:2201.06202, 2022.

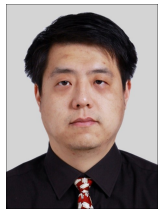
[168]XU J, XUE M, PICEK S. Explainability-based backdoor attacks against graph neural networks[C]// Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. 2021:31-36.

[169]YING R, BOURGEOIS D, YOU J, et al. Gnnexplainer: Generating explanations for graph neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32:9240.

[170]HUANG Q, YAMADA M, TIAN Y, et al. Graphlime: Local in-

terpretable model explanations for graph neural networks[J]. arXiv:2001.06216,2020.
[171]KIOURTI P,WARDEGA K,JHA S,et al. Trojdrl:Trojan attacks on deep reinforcement learning agents[J]. arXiv:1903.06638,2019.
[172]ASHCRAFT C,KARRA K. Poisoning Deep Reinforcement Learning Agents with In-Distribution Triggers[J]. arXiv:2106.07798,2021.
[173]WANG Y,SARKAR E,LI W,et al. Stop-and-go:Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems[J]. arXiv:2003.07859,2020.
[174]KRAJZEWICZ D,ERDMANN J,BEHRISCH M,et al. Recent development and applications of SUMO-Simulation of Urban MObility[J]. International Journal on Advances in Systems and Measurements,2012,5(3/4):128-138.
[175]GUO J,POTKONJAK M. Watermarking deep neural networks for embedded systems[C]// 2018 IEEE/ACM International Conference on Computer-Aided Design(ICCAD). 2018:1-8.
[176]ZHANG J,GU Z,JANG J,et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018:159-172.
[177]CAO X,JIA J,GONG N Z. IPGuard:Protecting intellectual property of deep neural networks via fingerprinting the classification boundary[C]// Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. 2021:14-25.
[178]LUKAS N,ZHANG Y,KERSCHBAUM F. Deep neural network fingerprinting by conferrable adversarial examples[J]. arXiv:1912.00888,2019.
[179]ADI Y,BAUM C,CISSE M,et al. Turning your weakness into a strength:Watermarking deep neural networks by backdooring[C]// 27th {USENIX} Security Symposium. 2018:1615-1631.
[180]XU J,PICEK S. Watermarking Graph Neural Networks based on Backdoor Attacks[J]. arXiv:2110.11024,2021.
[181]LI G,XU G,QIU H,et al. A Novel Verifiable Fingerprinting Scheme for Generative Adversarial Networks[J]. arXiv:2106.11760,2021.
[182]LI Y,ZHANG Z,BAI J,et al. Open-sourced Dataset Protection via Backdoor Watermarking[J]. arXiv:2010.05821,2020.
[183]SOMMER D M,SONG L,WAGH S,et al. Towards probabilistic verification of machine unlearning[J]. arXiv:2003.04247,2020.
[184]SUN Z,DU X,SONG F,et al. CoProtector:Protect Open-Source Code against Unauthorized Training Usage with Data Poisoning[J]. arXiv:2110.12925,2021.
[185]HOWARD G D. GitHub Copilot:Copyright,Fair Use,Creativity,Transformativity,and Algorithms[EB/OL]. https://gavinhoward.com/uploads/copilot.pdf.
[186]LIN Y S,LEE W C,CELIK Z B. What do you see? evaluation of explainable artificial intelligence(xai) interpretability through neural backdoors[J]. arXiv:2009.10639,2020.
[187]SHAN S,WENGER E,WANG B,et al. Gotta Catch'Em All:Using Honeypots to Catch Adversarial Attacks on Neural Networks[C]// Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020:67-83.
[188]WU S,HE Q,ZHANG Y,et al. Debiasing Backdoor Attack:A Benign Application of Backdoor Attack in Eliminating Data Bias[J]. arXiv:2202.10582,2022.
[189]THORNTON C,HUTTER F,HOOS H H,et al. Auto-WEKA:Combined selection and hyperparameter optimization of classification algorithms[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013:847-855.
[190]ERICKSON N,MUELLER J,SHIRKOV A,et al. Autogluon-tabular:Robust and accurate automl for structured data[J]. arXiv:2003.06505,2020.

YING Zonghao, born in 1997, postgraduate, is a member of China Computer Federation. His main research interests include adversarial attack and backdoor attack.



WU Bin, born in 1980, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include network security and covert communication.

(责任编辑:何杨)