深度神经网络后门防御综述

江钦辉, 李默涵, 孙彦斌

广州大学 网络空间安全学院/网络空间先进技术研究院 广州 中国 510006

摘要 深度学习在各领域全面应用的同时, 在其训练阶段和推理阶段也面临着诸多安全威胁。神经网络后门攻击是一类典型的面向深度学习的攻击方式, 攻击者通过在训练阶段采用数据投毒、模型编辑或迁移学习等手段, 向深度神经网络模型中植入非法后门, 使得后门触发器在推理阶段出现时, 模型输出会按照攻击者的意图偏斜。这类攻击赋予攻击者在一定条件下操控模型输出的能力, 具有极强的隐蔽性和破坏性。因此, 有效防御神经网络后门攻击是保证智能化服务安全的重要任务之一, 也是智能化算法对抗研究的重要问题之一。本文从计算机视觉领域出发, 综述了面向深度神经网络后门攻击的防御技术。首先, 对神经网络后门攻击和防御的基础概念进行阐述, 分析了神经网络后门攻击的三种策略以及建立后门防御机制的阶段和位置。然后, 根据防御机制建立的不同阶段或位置, 将目前典型的后门防御方法分为数据集级、模型级、输入级和可认证鲁棒性防御四类。每一类方法进行了详细的分析和总结, 分析了各类方法的适用场景、建立阶段和研究现状。同时, 从防御的原理、手段和场景等角度对每一类涉及到的具体防御方法进行了综合比较。最后, 在上述分析的基础上, 从针对新型后门攻击的防御方法、其他领域后门防御方法、更通用的后门防御方法、和防御评价基准等角度对后门防御的未来研究方向进行了展望。

关键词 后门防御; 后门攻击; 人工智能安全; 神经网络; 深度学习

中图法分类号 TP18 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2024.07.03

A Survey on Defense against Deep Neural Network Backdoor Attack

JIANG Qinhui, LI Mohan, SUN Yanbin

Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China

Abstract While deep learning is widely applied in various applications, it also faces many security threats in its training and inference phases. The neural network backdoor attack is a typical type of deep learning-oriented attack. An attacker can implant an illegal backdoor into deep neural network model during the training phase by employing techniques such as data poisoning, model editing or transfer learning. When the corresponding backdoor trigger appears in the inference phase, the attacked model will give the wrong output according to the attacker's intention. This kind of attack endows the attacker with the ability to control the output of the model through the backdoor trigger, which is highly concealed and destructive. Therefore, effective defense against neural network backdoor attacks is one of the important tasks to ensure the security of intelligent services, and it is also one of the important issues of intelligent algorithm confrontation. In this paper, the defense techniques for deep neural network backdoor attacks are reviewed from the field of computer vision. First, the basic concepts of neural network backdoor attack and defense are explained. The main attack methods are summarized into three categories, and the reasonable positions, advantages and disadvantages of the corresponding defense mechanisms are outlined. Then, according to the different stages of the defense mechanism, the current typical backdoor defense methods are divided into four categories: dataset-level, model-level, input-level, and certifiable robust defense. The methods of each category are analyzed and summarized in detail according to their applicable scenarios, stages and research status. At the same time, a comprehensive comparison of the specific defense methods involved in each category is made from the perspectives of defense principles, means and scenarios. Finally, on the basis of the above analysis, the future research directions of backdoor defense are prospected from the perspectives of defense methods against new backdoor attacks, backdoor defense methods in other fields, more general backdoor defense methods, and defense evaluation benchmarks.

Key words backdoor defense; backdoor attack; artificial intelligence security; neural network; deep learning

通讯作者: 李默涵, 博士, 教授, Email: limohan@gzhu.edu.cn。

本课题得到国家自然科学基金(No. 62372126, No. 62072130)、广东省自然科学基金面上项目(No. 2021A1515012307, No. 2020A1515010450)、广州市科技计划一般项目(No. 202102021207, No. 202102020867)、广东省高校创新团队项目(No. 2020KCXTD007)、广州市高校创新团队项目(No. 202032854)、广东省珠江学者岗位计划(2019)资助。

收稿日期: 2022-08-09; 修改日期: 2022-11-22; 定稿日期: 2024-03-15

1 引言

随着近年来人工智能技术的飞速发展, 深度神经网络 (Deep Neural Networks, DNN)在计算机视觉[1-2]、语音识别[3]和自然语言处理[4]等领域取得了空前卓越的成果。基于深度神经网络的应用系统逐渐出现在人们的日常生活中, 例如随处可见的人脸识别系统和新型智能汽车的自动驾驶辅助系统等。然而, 深度神经网络也被证明可能隐藏着巨大的安全威胁[5-6]。在训练阶段, 数据投毒[7] (Data Poisoning)攻击可以通过污染训练数据使得模型倾斜或不可用。在推理阶段, 对抗样本[8] (Adversarial Examples)攻击可通过对原始样本的微小扰动生成具有欺骗性的样本从而绕过模型检测, 模型窃取[9] (Model Extraction)攻击则可通过有技巧地收集模型的输入和输出, 窃取模型隐私数据。现今, 上述三类攻击已得到了学术界和工业界较多的关注和研究, 但近年来, 另一种新的、具有极高的隐蔽性和危害性的攻击逐渐引起各界关注, 这类攻击被称作为神经网络后门攻击[10-15] (Backdoor Attack)或神经网络木马攻击 (Trojan Attack)。

神经网络后门攻击通过在训练阶段向神经网络模型中植入非法后门, 使得后门触发器在推理阶段出现时, 模型输出会按照攻击者的意图偏斜。具体地, 攻击者在训练阶段通过训练集投毒、模型参数编辑等方式向神经网络模型植入恶意的后门。后门与特定类型的触发器相关, 被植入后门的模型在推理阶段接收到带有触发器的输入样本时, 会将样本分类为攻击者预先设定的类别, 而在接收到不带有触发器的正常输入时则表现良好。这类攻击具有极高的隐蔽性和危害性, 并且在目前深度学习应用场景下, 已经具备了充分的发起条件。

从隐蔽性角度来看, 这类攻击将模型输出与特定后门触发器相关联, 只有当触发器出现时, 模型输出才按照攻击者意图偏斜, 而触发器通常易被忽略甚至肉眼不可见, 因此, 除非用户有机会能明确意识到触发器存在, 否则很难发现模型是否受到了恶意攻击。

从危害性角度分析, 由于后门只有在触发器出现的情况下才会被激活, 而对于其他任何情况模型几乎保持原有性能。这使得攻击者对于模型具有了隐蔽的操控能力, 进而可能在某些场景下引发突发性的、严重的安全事件。试想, 如果一辆在高速公路行驶的自动驾驶汽车的交通标牌识别系统受到此类攻击, 在触发器突然出现时, 系统会将“限速”误判为“停车”, 就可能造成严重的交通事故。

在攻击发起条件方面, 当前的许多应用场景为神经网络后门攻击的孳生提供了土壤。昂贵的计算资源和缺乏高质量的数据是许多深度神经网络使用者面临的现实问题。因此, 使用第三方机器学习云平台, 例如机器学习即服务[16] (Machine Learning as a Service, MLaaS)或直接从互联网上下载机器学习模型和数据集成为了许多人的首选, 这为后门攻击创造了现实条件。恶意攻击者可以将经过特殊处理的恶意数据集或预先训练好的恶意模型上传到网上供用户下载, 如果用户直接使用这些恶意数据或模型, 而不进行检测处理, 则会大大增加后门感染和发动的风险。

神经网络后门攻击的隐蔽性、危害性和已经具备的发动条件表明, 对其防御方法的研究已经迫在眉睫。目前, 后门防御研究逐渐受到国内外高度重视, 但据我们所知, 目前国内还没有详细的针对后门防御研究的相关综述。为此, 本文聚焦神经网络后门防御工作, 从计算机视觉领域出发, 对后门防御方法进行总结和分析。

本文围绕深度学习的全生命周期和模型受到后门攻击的基本情形对后门防御进行分类阐述。通常, 模型训练和应用需经过数据采集、模型训练、模型测试、模型部署等几个基本步骤。本文将采集数据(包括数据预处理)和训练模型的过程称为模型**训练时**, 将测试模型及模型部署等阶段视为模型**训练后**。训练时和训练后分别对应了用户从第三方获取训练数据和用户从第三方获取模型这两种神经网络后门攻击的发动条件。进一步地, 根据防御机制的建立阶段和关注重点, 将神经网络后门防御方法大致分为 4 类, 如图 1 所示。

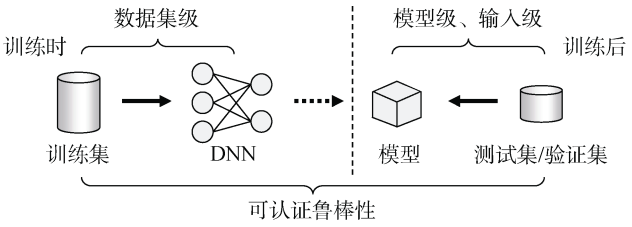

图 1 神经网络后门防御方法分类

Figure 1 Classification of neural network backdoor defense

其中, 模型训练时的防御关注数据集在训练过程中对模型的影响, 有对数据集本身进行改进的方法, 如检测异常数据、提供数据保障、增强训练数据等, 也有对训练方式进行改进的方法, 如改进训练

过程等, 两者的重点都是采取措施防止模型被污染的数据集蓄意破坏, 因此将其归纳为**数据集级别的防御**; 模型训练后的防御关注点大致可以分为两类, 一类关注模型本身, 如模型剪枝、微调、后门检测等, 因此将其归纳为**模型级别的防御**, 另一类关注触发器输入, 如检测触发输入、修复输入、输入预处理等, 因此将其归纳为**输入级别的防御**; 此外, 还有一类防御关注深度学习全生命周期且可提供坚实的理论保障, 本文将其归类为**可认证鲁棒性防御**。

在上述分类的基础上, 本文从如下方面梳理了神经网络后门攻击防御的相关研究:

(1) 总结了神经网络后门攻击和防御的相关概念定义;

(2) 综合考虑深度学习的全生命周期和模型受到后门攻击的现实场景, 将现有主流防御方法归纳为数据集级、模型级、输入级和可认证鲁棒性防御 4 类;

(3) 详细介绍了每类方法的基本原理和实现过程, 并从防御手段、防御场景、防御效果等角度系统地进行了分析比较;

(4) 基于神经网络后门防御工作的研究现状, 对未来方向进行了分析和展望。

2 神经网络后门攻击和防御的基础概念

DNN 由一系列参数和函数构成, 其通过拟合大量的训练数据确定数以万计的权重和偏置等参数。DNN 模型可表示为 $F(\boldsymbol{\theta},\boldsymbol{x})=y$, 其中 $\boldsymbol{\theta}$ 表示模型参数, $\boldsymbol{x}$ 表示待预测输入, y 表示模型对 $\boldsymbol{x}$ 输出的分类标签。

2.1 神经网络后门攻击

神经网络后门攻击的目的是向 DNN 植入隐藏后门, 使得后门触发器在推理阶段出现时, 模型输出按照攻击者的意图偏斜。神经网络后门攻击过程如下。令 $F(\cdot)$ 表示使用一组干净数据 $\boldsymbol{X}$ 训练得到的良性模型, $F'(\cdot)$ 表示基于后门攻击算法产生的后门模型, $F'(\cdot)$ 与预先设定的后门**触发器** (Trigger) $\boldsymbol{\delta}$ 关联。触发器是激活模型后门的钥匙。对于任意正常的**干净输入** $\boldsymbol{x}$, 都可以通过将触发器附着于 $\boldsymbol{x}$ 上得到**触发输入** $\boldsymbol{x}_{trigger}=\boldsymbol{x}+\boldsymbol{\delta}$, $F'(\cdot)$ 在推理阶段对于 $\boldsymbol{x}$ 和 $\boldsymbol{x}_{trigger}$ 有不同的行为表现: 对于 $\boldsymbol{x}$, $F'(\cdot)$ 保持与 $F(\cdot)$ 一样的输出, 即 $F'(\boldsymbol{x})=F(\boldsymbol{x})$; 对于 $\boldsymbol{x}_{trigger}$, $F'(\cdot)$ 会将其分类为攻击者预先设定的类别, 即 $F'(\boldsymbol{x}_{trigger})=y_{target}$, 其中 y_{target} 称为**目标标签** (Target Label)。

为产生 $F'(\cdot)$, 攻击者通常有三种策略, 即数据投毒、模型编辑、迁移学习, 如图 2 所示。用户从第三方获取训练数据时, 可基于数据投毒植入后门。用户从第三方获取模型时, 则可选用三种攻击的任意一种发起攻击。

策略 1 (数据投毒)。攻击者对训练数据进行操纵, 将部分 $\boldsymbol{x}\in\boldsymbol{X}$ 处理为 $\boldsymbol{x}_{poison}$, $\boldsymbol{x}_{poison}$ 由向 $\boldsymbol{x}$ 添加触发器 $\boldsymbol{\delta}$ 并打上目标标签 y_{target} 制作而成, 得到中毒数据集 $\boldsymbol{X}'$ ($\boldsymbol{x}_{poison}\in\boldsymbol{X}'$), 使用 $\boldsymbol{X}+\boldsymbol{X}'$ 训练神经网络, 得到模型 $F'(\cdot)$。

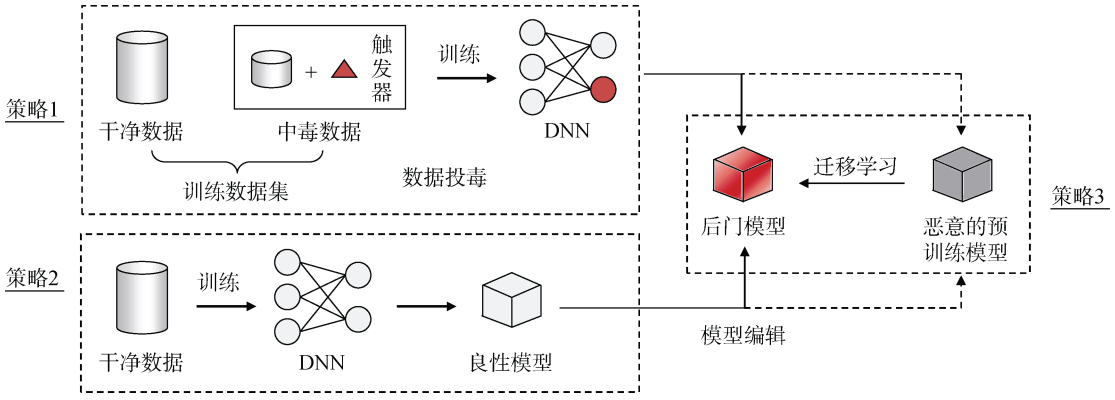

图 2 神经网络后门攻击过程

Figure 2 Backdoor attack processes on neural network

在计算机视觉领域, 触发器 $\boldsymbol{\delta}$ 通常是一块可见的像素图案, 如特定形状的像素点, 或特定的背景图片等。数据投毒是实施后门攻击的主要策略[10-12], 通过构造带有触发器 $\boldsymbol{\delta}$ 和目标标签的恶意数据并将其混入训练数据中, 使得 DNN 在模型训练时“记住”触发器, 即通过让触发器与特定神经元关联进而与目标标签形成强关联。从而, 在推理阶段, 一旦触发器出现在待分类输入样本上, 模型就会以很大概率

将该输入识别为目标标签。

策略 2 (模型编辑)。攻击者对模型进行编辑, 通过修改模型权重参数直接操纵模型产生后门, 使得 $F(\cdot)$ 转化为 $F'(\cdot)$。

模型编辑使攻击者能够在不操纵训练数据集的情况下向 DNN 植入后门[13]。通过探索模型中神经元对输入数据的敏感性, 可以发现候选后门神经元, 进而利用特殊构造的触发器, 指导模型参数调整, 实现神经网络后门植入。

策略 3 (迁移学习)。攻击者先基于数据投毒或模型编辑生成恶意教师模型, 再通过迁移学习引导受害学生模型产生后门。

一些典型的迁移学习方法通过对预先训练好的模型进行二次训练来实现知识迁移, 并生成新的神经网络模型。后门攻击者可以利用恶意的预训练模型传播后门[14], 诱导迁移后的模型被植入后门。

2.2 神经网络后门防御

根据深度学习的生命周期来划分, 神经网络后门攻击的防御机制可以建立于不同的阶段或位置。首先, 防御机制可以建立于模型训练时, 直接从训练数据入手, 通过检测异常数据、增强训练数据、改进训练过程等手段, 来达成防御目的; 其次, 防御可以建立在模型训练后, 一方面, 可以直接从模型本身入手, 通过模型剪枝、微调、后门检测等手段, 发现和修复后门, 另一方面, 也可以聚焦于潜在的触发输入, 通过输入预处理、检测触发输入、修复输入等手段, 防止模型接收到恶意输入; 此外, 也可以令防御机制覆盖深度学习的全生命周期, 通过证明模型鲁棒性确保模型的训练和推理阶段不会被后门攻击影响。

综上, 本文根据防御机制建立的阶段或位置, 从数据集级、模型级、输入级和可认证鲁棒性四个角度讨论神经网络后门防御。

数据集级防御机制建立于模型训练时, 主要关注训练数据的安全性及其对模型的影响, 通过在训练阶段采取防御措施来降低模型受到后门攻击的风险。主要的防御措施包括检测训练集中是否存在中毒数据、和对训练数据或训练过程进行优化, 从而抑制中毒数据防止模型产生后门。

模型级防御机制建立于模型训练后, 主要从模型本身出发, 尝试检测或修复后门模型。其中, 后门模型检测方法种类较多, 但大部分方法的基本思想可以归结为两大类: 一是通过逆向工程来恢复后门触发器, 旨在寻找触发器来识别模型是否存在后门; 二是对大量良性模型与后门模型进行特征提取, 基于提取到的特征训练新的分类器, 以预测目标模型是否存在后门。后门模型修复方法则主要通过剪枝、微调、模型优化等手段对潜在的被植入后门的神经网络进行调整, 尝试清除后门。

输入级防御机制同样建立于模型训练后, 重点关注模型进行推理时输入数据的安全性。后门攻击导致的恶意行为只有在触发输入出现时生效, 因此检测并拦截恶意的触发输入是有效防御后门攻击的一种方式, 此外还可以通过输入预处理来减少模型后门被激活的可能性。

可认证鲁棒性防御是一类覆盖深度学习全生命周期的防御机制, 其通过验证模型的鲁棒性来提供安全保障。在一定条件下, 可以保证模型在某种程度或范围内免疫后门攻击。这种防御从理论层面保证了条件满足时防御必然成功, 因此可以有效缓解攻击者和防御者之间的“军备竞赛”, 但目前相关研究还相对匮乏。

此外, 还有一些防御手段可以通过拼接不同阶段的防御机制来完成全生命周期防御, 这类手段很大程度上与单阶段的防御存在重叠, 因此本文不再赘述这类拼接的防御方法。

3 数据集级防御方法

数据集级防御的应用阶段为模型训练时, 关注防御方从第三方获取训练数据的场景。这类场景下, 潜在的攻击者可能会采用投毒的方式, 将包含触发器的中毒样本混在训练数据中提供给防御方, 防御方如果直接使用包含中毒样本的数据进行训练, 则可能得到后门模型。因此, 防御方有必要在模型训练时采取必要的措施, 检测、隔离或抑制中毒样本在训练过程中所起的作用, 从而发现和预防后门攻击。表 1 按时间顺序总结对比了主要的数据集级防御方法的原理及防御手段。

3.1 方法介绍

谱特征 (Spectral Signatures)。神经网络在训练过程中挖掘隐藏在训练数据中的特征。由于后门攻击可以改变神经网络在触发输入上的预测, 因此被攻击的网络包含比正常特征更强的后门相关特征。Tran 等人[17]表示, 后门攻击易在神经网络学习表征的协方差谱中留下可被检测的痕迹, 称之为谱特征。他们提出了一种基于主成分分析 (Principal Component Analysis, PCA)的防御算法。首先使用包含中毒数据的训练集训练网络, 并提取所有训练数据的中间层表征。然后对表征的协方差矩阵进行奇异值分解 (Singular Value Decomposition, SVD), 为每一个

表 1 数据集级防御方法

Table 1 Defense at data set level

方法	时间	原理/简述	防御手段	防御场景	防御效果
谱特征[17-18]	2018	计算输入数据中间层表征的离群分数	检测异常数据	从第三方获取训练数据	可有效检测常见的像素攻击, 后续研究对谱特征不显著的攻击也能适应
激活聚类[19]	2019	对输入数据的激活值进行聚类和分析	检测异常数据	从第三方获取训练数据	可有效检出文本和图像数据集上的简单攻击, 对复杂中毒方案也有一定效果
提取中毒信号[20]	2019	从输入数据的梯度中提取后门中毒信号	检测异常数据	从第三方获取训练数据	可检测常见像素攻击和图像混合攻击
梯度塑造[21]	2020	对训练数据的梯度进行裁剪和扰动	提供数据保障	从第三方获取训练数据	可减少清洁标签攻击一半的攻击成功率
数据增强[22-23]	2021	有针对性的添加可以减弱后门攻击的训练样本	增强训练数据	从第三方获取训练数据	对普通像素攻击, 毒害所有目标类图像, 可显著降低攻击成功率, 并提升少量模型性能 对普通像素攻击, 可有效降低攻击成功率, 迁移学习时也有一定效果
频域分析[24]	2021	将训练数据转换至频域进行研究和分析	检测异常数据	从第三方获取训练数据	对多种不同触发器的攻击, 达到 98.50% 的平均检出率
反后门学习[25-26]	2021—2022	在毒化训练集上训练出干净的模型	改进训练过程	从第三方获取训练数据	可对常见的脏标签攻击、清洁标签攻击和特征空间攻击进行防御 可有效防御普通像素攻击、图像混合攻击和 WaNet 攻击

训练数据计算离群分数。基于离群分数对训练数据集进行过滤, 剔除分数过高的数据。最后, 在过滤后的数据上训练新模型。

最近, Hayase 等人[18]提出可以利用鲁棒协方差估计器 (Robust Covariance Estimator)来获得干净数据特征的近似均值和协方差, 并用其对训练数据进行白化处理, 从而放大隐藏的谱特征。此外, 还引入了 QUE (QUantum Entropy)异常评分体系来增强防御的鲁棒性。

激活聚类 (Activation Clustering)。Chen 等人[19]通过分析训练数据的神经元激活情况确定目标网络是否受到攻击。他们认为末位隐藏层的激活情况更适用于探测中毒信号, 而靠前的隐藏层则会给后续分析增加噪声。该方法流程大致如下。首先, 在可能包含中毒数据的不可信数据集上训练神经网络, 然后利用训练数据对网络进行查询, 得到网络中最后一层隐藏层的激活情况。接着, 使用独立成分分析 (Independent Component Analysis, ICA)对训练数据的激活进行降维, 对每一类标签执行 k-means 聚类算法 ($k=2$), 将该类别下的训练数据聚为两簇, 记为 C_1 和 C_2。将 C_1 和 C_2 分别从训练数据集中剔除, 重新训练得到两个新模型, 记为 M_1 和 M_2。然后, 使用 C_1 作为测试数据去测试 M_1, C_2 作为测试数据去测试 M_2, 如果 C_1 和 C_2 中均不存在中毒的训练数据, 则 M_1 和 M_2 的行为应当是类似的, 即均能正确预测 C_1 和 C_2 中的大部分数据, 反之, 则可能存在包含触发器的中毒训练数据。此外, 作者还发现干净数据对应的激活倾向于被分成两个大小相近的簇, 如果存在中毒数据对应的激活, 则其往往处于一个较小的簇中。

提取中毒信号。Chan 等人[20]提出, 可以从输入层的梯度中提取关于损失函数的中毒信号, 并利用该信号过滤训练集中的中毒数据并检测目标标签。在被感染的后门模型中, 存在中毒的神经元, 其在触发输入出现时, 激活值会远高于其他正常的神经元, 从而可以主导分类器预测。中毒神经元的权重的绝对值高于其他神经元, 而数据的输入梯度线性依赖于激活函数的导数与权重。因此, 触发器所在的特殊像素位置应具有较大的绝对输入梯度值。基于此结论, 作者设计了一种允许梯度噪声存在的 SVD 方法实现中毒信号的提取。根据训练数据输入梯度和中毒信号的余弦相似度对干净数据和中毒数据进行聚类, 能够过滤出所有的中毒数据并发现目标标签, 而通过观察中毒数据的梯度分量, 也可以找到正确的原始标签。

梯度塑造 (Gradient Shaping)。Hong 等人[21]发现中毒数据梯度的 L2 范数相比干净数据的梯度拥有更高的量级, 其梯度方向也与干净数据不同。因此, 可以通过限制训练数据的梯度大小和梯度方向上的差异来进行防御。该方法通过在训练过程中裁剪和扰动梯度, 对梯度进行塑造, 在保证梯度统计量基本准确的前提下, 削弱单个训练样本对整体训练过程的影响。梯度塑造提供了应对后门数据投毒攻击的一种通用防御思路, 它不依赖于自适应触发器模式, 可以作为模型防御后门的一个重要环节, 在训

练过程中从梯度层面削弱后门攻击的强度。

数据增强。Borgnia 等人[22]通过研究发现, 或许可以基于数据增强技术在不牺牲模型性能的前提下降低后门攻击的成功率。他们在 CIFAR-10 数据集上设置了两个简单的后门攻击任务, 验证了数据增强在防御后门攻击上的有效性。在一些条件下, 攻击成功率可以从 100%下降为 36%, 而同时验证集准确率提升了 9%。此外, Geiping 等人[23]提出, 可以将对抗样本攻击防御中使用的数据增强思想扩展到后门攻击防御中, 通过有策略地制作“毒药”并将其注入到训练过程中, 可以降低神经网络对中毒数据引起的扰动的敏感性。基于数据增强的防御方法研究现在尚不充分, 但可以预见, 数据增强的方法应当是一类相对有效的防御方法。这是因为, 基于数据投毒的后门攻击的目标是尝试通过篡改训练数据来使得训练数据分布偏斜, 进而建立触发器和目标标签的强关联, 而有针对性的数据增强可以修正这种不合理偏斜以达成防御目标。

频域分析。频域 (Frequency Domain)常用于描述信号在频率方面的特性, Zeng 等人[24]提出从频域的角度分析神经网络后门攻击。他们通过对主流后门触发器的重新思考和分析, 指出当前常见的触发器模式往往在频域产生明显的人工痕迹, 称之为高频伪影 (High-frequency Artifacts)。作者利用离散余弦变换 (Discrete Cosine Transform, DCT)将训练集中的图像转换至频域, 并将 DCT 频谱绘制为易于观察的热力图。他们对多种触发器类型进行了研究和分析, 发现正常图像的频谱主要包含低频分量, 而具有触发器的恶意图像往往包含更多的高频分量。利用这一结论, 作者设计了一种基于频域分析的后门图像检测方法。

反后门学习。Li 等人[25]提出, 可以直接在中毒数据上尝试学习干净的良性模型, 这种学习方法称为反后门学习。他们将神经网络模型在毒化训练集上的训练任务视为两个子任务, 一是让模型学习干净数据的清洁任务, 二是让模型学习中毒数据的后门任务。通过对 10 种目前主流的后门攻击的研究发现, 在模型训练的过程中, 后门任务要比清洁任务更容易地完成, 具体表现为包含触发器的中毒数据的平均损失在模型训练早期远小于干净数据的平均损失。因此, 可以在训练早期通过“局部梯度上升”检出部分中毒数据并进行隔离, 在训练后期, 采用“全局梯度上升”遗忘在早期阶段模型学到的后门知识, 以此得到干净的良性模型。

此外, Huang 等人[26]提出的基于解耦的防御方法也可以归类为一种反后门学习。他们发现在标准监督学习的训练程序下, 中毒数据的隐藏特征形成了单独的聚类, 而在对去除标签的毒化训练集进行自监督训练后, 中毒数据则和对应的干净数据有着相似的隐藏特征。因此, 可以通过训练程序解耦防御后门攻击, 即, 首先在去除标签的毒化训练集上进行自监督学习得到特征提取器, 接着在完整的毒化训练集上进行标准监督学习, 训练剩余的全连接层。由于特征提取器与标签无关, 所以中毒数据和干净数据在隐藏特征空间中是相似的, 进而第二阶段标准监督训练相当于噪声标签学习[27], 可以通过半监督学习微调来得到良性模型。

3.2 小结

综上所述, 数据集级防御研究异常训练数据与神经网络后门之间的强关联性来检测网络是否存在后门。一些方法利用神经网络中间层特征的统计信息来发现中毒数据, 通过隔离或弱化中毒数据来防止后门, 如谱特征[17-18]、激活聚类[19]、频域分析[24]等。这些方法通常简单易实现, 但对数据本身较为依赖, 两类数据激活差异的不稳定性限制了防御鲁棒性的提升。还有一些方法在数据训练过程中发现中毒数据和干净数据之间的行为差异, 如梯度和损失差异。可以检测到中毒数据, 并通过纠正这些差异来防止后门, 例如提取中毒信号[20]、梯度塑造[21]、反后门学习中基于梯度上升的方法[25]等。相比之下, 这些方法对中毒数据的独特学习行为更加敏感, 因此通常可以适应更多的触发器类型, 但防御的训练过程相对复杂。此外, 在很多场景下用户仅从第三方获得模型, 而无法访问训练数据, 此时数据集级的防御方法难以应用。

4 模型级防御方法

模型级防御的应用阶段为模型训练后, 关注防御方从第三方获取模型的场景。这类场景下, 防御者无法访问模型初始训练过程, 仅针对训练好的潜在受害模型进行后门检测和修复。可以将这类防御方法进一步分成两大阶段, 第一阶段对给定模型检测其是否存在后门, 这阶段的防御方法大致可以分为基于逆向触发器的防御方法和基于元分类器的防御方法, 第二阶段对可能存在后门的模型进行修复, 具体可采用剪枝、微调、模型优化等手段。表 2 按时间顺序分类总结对比了主要的模型级防御方法的原理及防御手段。

4.1 方法介绍

下面将依次介绍逆向触发器、元分类器、模型修复和其他类型的模型级防御方法。

表 2 模型级防御方法

Table 2 Defense at model level

分类	方法	时间	原理/简述	防御手段	防御场景	防御效果
逆向触发器	神经清洁[28]	2019	首个构建逆向触发器防御后门的工作	检测后门模型	从第三方获取模型	可逆向小尺寸触发器, 有效检测和防御 BadNets 攻击和 Trojan 攻击
	Tabor[29]	2019	重新设计损失函数的正则项, 优化搜索过程	检测后门模型	从第三方获取模型	可逆向较大尺寸的触发器
	ABS[30]	2019	刺激神经元, 优化逆向触发器搜索过程	检测后门模型	从第三方获取模型	可逆向不同尺寸和形状的触发器, 在少量干净样本的条件下, 可有效检出后门
	MESA[35]	2019	利用阶梯逼近恢复泛化的原始触发器	检测后门模型	从第三方获取模型	可生成小尺寸触发器上的有效触发分布, 提高防御鲁棒性
	DeepInspect[33]	2019	基于 CGAN 构建逆向触发器	检测后门模型	从第三方获取模型	对比神经清洁, 不依赖干净样本, 检测精度和速度均有一定提升
	GanSweep[32]	2020	基于 GAN 构建逆向触发器	检测后门模型	从第三方获取模型	不依赖干净样本, 可有效防御大尺寸、可变和多触发器的攻击
	NNoculation[34]	2020	基于 Cycle-GAN 构建逆向触发器	检测后门模型	从第三方获取模型	可有效检测 BadNets 攻击, 对比神经清洁和 ABS, 性能有一定提升
	K-Arm[31]	2021	通过迭代与选择策略优化逆向触发器搜索过程	检测后门模型	从第三方获取模型	对比神经清洁、Tabor 和 ABS, 大大优化了逆向触发器搜索时间与检测精度
	B3D[37]	2021	采用无梯度优化算法构建逆向触发器	检测后门模型	从第三方获取模型	可检测黑盒场景下的后门, 对比神经清洁和 Tabor, 检测精度有一定提升
	像素优化[36]	2022	通过最小化像素变化构建逆向触发器	检测后门模型	从第三方获取模型	可逆向出扰动像素更少、攻击成功率更高、鲁棒性更强的触发器
元分类器	通用石蕊模式[38]	2020	为模型添加石蕊模式, 训练元分类器	检测后门模型	从第三方获取模型	可有效检测常见的像素攻击, 对比神经清洁, 检测 AUC 有所提升
	单像素签名[39]	2020	提取模型的特殊签名, 训练元分类器	检测后门模型	从第三方获取模型	训练与测试的模型不同时, 也有较高的的后门检出率, 对比神经清洁和 ABS, 平均检出率有明显提升
	元神经木马检测[40]	2021	优化木马设置与查询集, 训练元分类器	检测后门模型	从第三方获取模型	可有效检测图像、语音、文本等数据集上的常见木马攻击, 检测 AUC 优于激活聚类、谱特征、神经清洁等方法
模型修复	精细剪枝[41]	2018	修剪休眠神经元, 微调模型	剪枝、微调	从第三方获取模型	相比从零开始训练, 精细剪枝更有效, 可有效修复普通的后门模型
	模式连接修复[42]	2020	利用模式连接在损失路径中选择最优模型	模型优化	从第三方获取模型	可有效降低常见攻击的对抗性效应, 保持一定的模型性能
	神经注意力蒸馏[43]	2020	利用教师模型引导学生模型忘记后门	知识蒸馏	从第三方获取模型	在少量干净样本条件下, 可有效消除 Trojan、SIG、Refool 等 6 种攻击产生的后门, 优于精细剪枝和模式连接修复
	BDA[44]	2021	提取输入数据对应的关键神经元, 定位异常神经元	剪枝	从第三方获取模型	可有效识别和移除单目标攻击和多目标攻击产生的后门, 对比神经清洁中的剪枝, 具有更低的神经元修剪率
	对抗神经元剪枝[45]	2021	利用神经元扰动的敏感性定位后门相关神经元	剪枝	从第三方获取模型	在少量干净样本条件下, 可有效修复多种不同的后门, 对比精细剪枝和模式连接修复, 模型性能下降的更少
	CARE[46]	2021	基于因果推断对故障神经元进行定位和修复	故障定位	从第三方获取模型	可有效修复常见的后门模型, 相比传统的再训练和微调有更少的时间开销
	AI-Lancet[47]	2022	基于特征分析对故障神经元进行定位和修复	故障定位	从第三方获取模型	可有效修复 BadNets、Trojan 等攻击产生的后门, 优于神经清洁
	I-BAU[48]	2022	提出后门修复 minimax 优化问题, 并进行求解	模型优化	从第三方获取模型	对比 DeepInspect、Tabor、精细剪枝、NAD 等方法, I-BAU 对触发器、中毒预算等攻击设置有更强的鲁棒性和性能
其他	NeuronInspect[49]	2019	为每类标签生成解释热力图, 寻找后门	检测后门模型	从第三方获取模型	可有效检测常见像素攻击、多触发器和半透明触发器的攻击, 对比神经清洁, 在鲁棒性和效率方面有明显提升
	DL/DF-TND[50]	2020	数据有限和无数据条件下的后门检测	检测后门模型	从第三方获取模型	在缺少训练样本条件下, 可有效检出多种形状、颜色和位置触发器产生的后门

4.1.1 逆向触发器

基于逆向触发器的后门攻击检测及防御方法是目前最先进的方法之一。防御者通过搜索触发器或借助生成模型 (Generative Model)逆向生成攻击者预置的触发器来检测后门。

有相当多的方法通过搜索触发器来逆向寻找可能的触发器模式。Wang 等人[28]率先提出了逆向触发器的概念和基于逆向触发器的后门防御方法, 称为神经清洁 (Neural Cleanse)。首先, 对样本中的所有标签, 依次视其为后门攻击的潜在目标标签, 对于每个潜在目标标签 t, 搜索将其他标签的样本误分类为 t 所需的最小扰动 $\boldsymbol{\delta}_t$, 将 $\boldsymbol{\delta}_t$ 作为候选逆向触发器。然后, 通过离群点检测, 挑出异常小的候选逆向触发器作为最终选取的逆向触发器。可以基于最终选取的逆向触发器进一步检测触发输入, 或寻找与触发器相关的异常神经元。

随后, 越来越多的科研工作者开始关注这类防御方法。由于该方法的有效性很大程度上取决于逆向触发器的质量, 因此一些后续研究工作致力于提出更优的搜索策略。

一些方法在神经清洁的基础上优化搜索过程。Guo 等人[29]提出的 Tabor 和神经清洁[28]一样, 都将后门检测视为一个优化问题, 即在对抗子空间 (Adversarial Subspace)中搜索触发输入。作者改进了神经清洁中搜索逆向触发器的目标函数, 通过设计新的正则项减少搜索过程中遇到的无关对抗性样本。Liu 等人[30]通过分析 DNN 内部神经元的异常行为, 优化逆向触发器的生成过程。Shen 等人[31]受强化学习中多臂老虎机 (Multi-arm Bandit)问题的启发, 提出了一种用于后门模型检测的 K-Arm 优化算法。在目标函数的指导下, 通过迭代与随机选择最有希望的标签对逆向触发器的搜索过程进行优化。

与此同时, 一些工作提出可以利用生成模型来构建逆向触发器。Zhu 等人[32]利用生成对抗网络 (Generative Adversarial Networks, GAN)构建逆向触发器。Chen 等人[33]提出的 DeepInspect 框架首先利用模型反演来恢复初始训练集, 然后使用条件生成对抗网络 (Conditional Generative Adversarial Networks, CGAN)构建逆向触发器。Veldanda 等人[34]首先使用随机扰动后的验证集对 DNN 模型进行再训练, 得到具有一定“抗后门”能力的微调模型。同时, 将微调模型与原模型拥有不同输出的输入样本进行隔离, 得到“隔离数据集”。接着, 利用循环生成对抗网络 (Cycle-GAN)学习干净数据和隔离数据之间的转换 (即逆向触发器), 构建可能的拥有清洁标签的触发输入数据集, 使用该数据集对目标模型进行再训练可以缓解后门。

此外, 还有一些工作从其他角度试图寻找更好的搜索策略。Qiao 等人[35]发现模型在被后门注入的过程中泛化了原始的触发器, 因此他们使用基于最大熵的阶梯近似法 (Max-Entropy Staircase Approximator, MESA)恢复了连续的触发分布, 从而达到相比于基线更好的防御效果。最近, Tao 等人[36]设计了一种可以直接对每一个独立像素进行扰动优化的触发器逆向算法。相较于神经清洁[28], 该算法不需要使用掩码 (Mask), 生成的逆向触发器拥有更少的像素扰动和更高的攻击准确率。Dong 等人[37]研究了关于生成逆向触发器的黑盒策略。作者将搜索逆向触发器的优化过程视为黑盒模型, 提出一种基于自然进化策略 (Natural Evolution Strategies)的优化算法。

4.1.2 元分类器

元分类器 (Meta-classifier)是“用来对分类器进行分类的分类器”。最近, 元分类器的思想也被用来检测模型是否存在后门。一些工作将模型级后门检测任务视为这样一个二分类任务: 给定任意 DNN 模型 m 作为输入, 元分类器输出 0 或 1, 0 表示 m 是良性的, 1 表示 m 存在后门。元分类器防御的工作流程如图 3 所示。

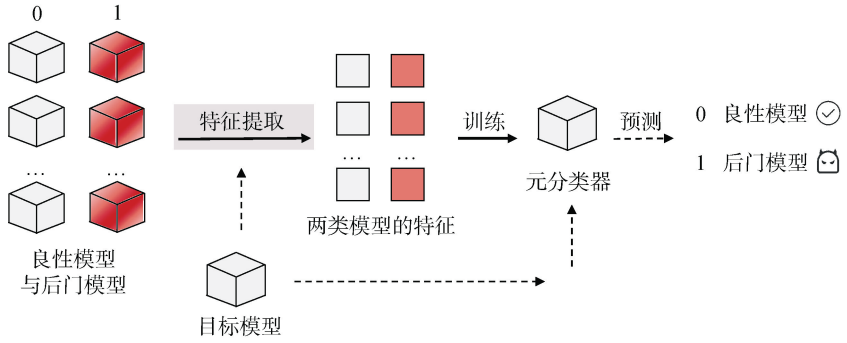

图 3 元分类器防御工作流程

Figure 3 The workflow of meta-classifier defense

这类防御方法的主要时间开销在于元分类器训练, 一旦元分类器训练完成, 其诊断速度是非常快的, 因为只需要通过几次简单的神经网络正向传播就可以判断模型中是否存在后门。这类方法的研究目前还不多见, 但若可以有效构建元分类器, 则有可能实现更为高效、易于部署的后门检测。

Kolouri 等人[38]提出通用石蕊模式 (Universal Litmus Patterns, ULPs)的概念, 他们认为, 可以通过生成一组 ULPs, 分析网络对这些模式的输出来揭示后门攻击。作者的思路是对每一个数据集训练数百个 NN (Neural Network)模型, 作为训练元分类器的数据集, 这些 NN 模型包含了良性模型和后门模型。具体做法是将一组 ULPs 嵌入单个 NN 模型中, 然后拼接模型的 Logit 输出, 作为元分类器的输入特征。训练元分类器时, 一同优化和更新 ULPs, 从而得到最优的 ULPs 和元分类器。

Huang 等人[39]定义了一种单像素签名 (One-pixel Signature), 它是输入图像中对目标模型预测值变化影响最大的像素的集合。与 ULPs 方法[38]相似, 作者同样使用大量良性和后门 NN 模型训练元分类器。特别之处在于, 作者对给定的 NN 模型提取签名作为元分类器的输入特征, 这些签名能够很好的揭示良性模型和后门模型之间的差异。此外, 作者还根据已有攻击的触发器模式生成通用的“疫苗”模式, 从而创建更加可靠的后门模型用于训练, 增强元分类器鲁棒性。

单像素签名[39]方法中, 生成后门模型的“疫苗”模式是基于 BadNets[10]攻击方法生成的, 其只考虑了不同颜色和位置的多边形触发器模式 (Polygon Trigger Patterns)。Xu 等人[40]提出元神经木马检测 (Meta Neural Trojan Detection)方法, 设计了一种生成方法用于生成更丰富的后门类型, 囊括像素攻击和图像混合攻击在内的多种触发器模式, 通过巨量学习 (Jumbo Learning)来训练后门影子模型 (Shadow Model), 这些后门影子模型和良性模型一起被用于训练元分类器。为了训练出一个高性能的元分类器, 作者提出可以使用一组优化查询集更好地提取 NN 模型特征, 该查询集揭示了 NN 模型在重要输入上的行为模式。

4.1.3 模型修复

模型修复的目的是缓解模型中潜在的后门, 修复受损网络。一些修复方法可以在后门检测之后使用, 有针对性地修复给定的后门模型, 另一些方法可以作为一种较为通用的安全性提升策略, 在不知道是否存在后门攻击的情况下提升模型安全性。

在这类工作中, 模型剪枝和微调提出最早, 相对来讲也最易于实现。剪枝原是一种模型压缩方法, 但通过修剪模型中后门相关的神经元则可以起到移除潜在后门的作用。微调或模型再训练是指使用部分数据对模型继续进行训练, 通过提升神经网络模型准确率来消除后门影响。早期, Liu 等人[51]在提出 Neural Trojans 攻击的同时, 指出可以通过模型再训练的方法修复后门模型, 他们用一组干净的数据继续训练被攻击的神经网络模型, 试图让模型忘记后门知识。Gu 等人[10]在提出 BadNets 攻击时表示, 触发输入可能触发了在干净输入时处于休眠状态的神经元。因此一个自然的直觉是, 若能调整网络, 将这些后门神经元移除, 就能够达到修复后门模型的效果。

Liu 等人[41]提出精细剪枝 (Fine-Pruning)方法。首先, 作者单纯使用剪枝策略, 尝试将一组干净数据输入被攻击的神经网络模型, 记录每个神经元的平均激活情况, 根据激活强度的递增顺序从小到大迭代修剪网络中的神经元。每一次修剪之后, 计算模型精度, 当精度低于一定阈值时停止修剪。这种方法在后门仅与休眠神经元关联的情况下可以一定程度上抵御后门攻击, 但是, 如果存在一种攻击, 让后门相关权重的学习由包含正常行为的神经元 (即非休眠神经元)来完成, 则上述策略可能效果不好。若强行修剪这种后门神经元, 会导致模型的精度大幅下降。接着, 作者单纯使用微调技术, 尝试利用一组干净数据对模型进行标准微调, 结果显示这种方法的效果也不尽如人意, 其原因可能是干净数据上的微调可能无法影响一些后门神经元。最后, 基于上述两项研究, 作者提出了精细剪枝方法, 将剪枝与微调进行结合, 先使用剪枝法尽可能的移除后门神经元, 然后使用微调法恢复部分因剪枝而降低的模型性能, 最终成功的消除了模型中的后门。

在一些干净输入上休眠的神经元未必总是异常神经元, 为了更准确地定位后门相关的异常神经元, Jiang 等人[44]提出了一种基于神经元剪枝的模型修复方法 BDA (Backdoor Defense Algorithm), 作者为寻找异常神经元进行精确的剪枝, 做了周密和详细的研究分析。该方法首先为 DNN 中的每个神经元部署控制门, 将模型的功能表示为神经元对输入数据语义的敏感性, 利用神经元的激活频率为每个输入数据提取一组关键神经元, 即对模型功能更有影响的神经元。然后, 根据关键神经元上的激活频率分布, 基于相关系数和异常指数两种指标定位异常神经元。异常神经元通常揭示了触发输入与原始类不相关、但与触发器强相关的特性。最后, 对 DNN 中的

每一层进行单独处理, 准确修剪每一个异常神经元。

所谓对抗性扰动, 是指附着于触发输入上的后门触发器, 其促使模型错误分类。Wu 等人[45]从对抗性扰动出发, 探究神经元上的直接扰动对模型预测结果的影响, 从而提出对抗神经元剪枝方法来修复后门模型。他们使用扰动因子对神经元权重和偏差进行扰动, 发现良性模型和后门模型在扰动后会出现不一致的行为表现。基于对抗性神经元扰动的敏感性, 可以定位后门相关的敏感神经元, 进而修剪这些敏感神经元修复后门模型。

除了剪枝和微调之外, 一些其他方法也被应用于后门模型的修复工作之中。

模式连接 (Mode Connectivity)通过训练两个模型之间的高精度路径揭示神经网络中损失函数的最佳值, 一些工作将其应用于模型优化[52]。Zhao 等人[42]采用模式连接技术在神经网络的损失曲面 (Loss Landscape)上研究 DNN 的对抗鲁棒性。对于给定的目标模型, 作者使用有限的干净数据训练出一个小的微调模型, 然后将此良性的微调模型与目标模型进行模式连接, 训练出一条高精度的损失路径, 从而选择一个鲁棒的 DNN 模型, 有效缓解了触发输入对模型的不良影响。在微调模型和后门模型连接的过程中, 与后门相关的神经元路径在 DNN 中被成功的消除, 从而达到了模型修复的目的。

Li 等人[43]提出了一种基于知识蒸馏[53]和注意力迁移的后门修复方法, 称为神经注意力蒸馏 (Neural Attention Distillation, NAD)。该方法首先在后门模型上使用一组干净数据进行标准微调来获得良性教师模型, 接着利用教师模型指导后门学生模型在一小部分干净训练数据上进行调整, 使得学生模型的中间层注意力与教师模型的注意力保持一致。具体来说, NAD 通过在 ResNet 模型的每个残差组之后计算注意力表征, 根据教师模型和学生模型的注意力表征定义 NAD 蒸馏损失, 引导后门学生模型忘记后门触发器。

最近, Zeng 等人[48]定义了神经网络后门模型修复的 minimax 优化问题, 如下所示:

$$\boldsymbol{\theta}^{*}=\arg\min_{\boldsymbol{\theta}}\max_{\|\boldsymbol{\delta}\|\leqslant C_{\boldsymbol{\delta}}} H(\delta,\ \theta):\ =\frac{1}{n}\sum_{i=1}^{n}L(f_{\theta}(\boldsymbol{x}_i+\boldsymbol{\delta}),\ \boldsymbol{y}_i)$$

其中, L 表示损失函数, $\boldsymbol{\theta}^{*}$ 和 $\boldsymbol{\theta}$ 是模型参数, $\boldsymbol{\delta}$ 表示触发扰动, $C_{\boldsymbol{\delta}}$ 是扰动预算。该式子通过集成内部最大化问题和外部最小化问题建立搜索触发扰动和搜索模型参数这两个优化问题的相互依赖性。内部最大化问题通过搜索触发器来寻找最坏情况下的攻击收益, 外部最小化问题通过调整模型参数来最小化最坏攻击收益。在此基础上, 作者设计了一种基于隐式超梯度 (Implicit Hypergradient)的 I-BAU (Implicit Backdoor Adversarial Unlearning)防御算法来更新后门模型, 只需利用少量干净数据就能修复受损的神经网络模型。

此外, 现有的一些基于故障分析和定位的神经网络修复方法在后门修复中也取得了不错的效果。

Zhao 等人[46]提出的 AI-Lancet 是一种基于神经元定位的模型修复方法。该方法先是找到误分类样本上的关键区域, 将移除此区域的样本与原样本的特征差异进行详细地分析, 再通过渐进式的消融法最终定位诱导错误的神经元。他们提出了两种基于已知故障神经元的修复方法, 一是直接反转该神经元值的符号, 消除其对模型输出的影响, 二是基于联合训练和知识蒸馏对模型进行神经元微调。

Sun 等人[47]则提出了一种基于因果关系的神经网络修复方法, 称为 CARE (CAusality-based REpair), 与 AI-Lancet[46]相似, 该方法也包括故障定位和精确修复两个步骤。在故障定位方面, CARE 将神经网络转化为因果模型, 基于定义的归因问题, 定位最可能诱发错误的神经元。在精确修复方面, CARE 采用粒子群优化 (Particle Swarm Optimisation)算法, 通过调整与故障神经元相关的模型参数来修复模型。

4.1.4 其他方法

除上述几类方法外, 还有一些从其他角度出发的模型级防御方法。

NeuronInspect。Huang 等人[49]提出了一种基于可视化解释技术的后门防御方法。作者使用一组干净的数据为 DNN 模型生成每个类别标签对应的激活热力图 (Heat Map), 观察和分析热力图的稀疏性、平滑性和持续性三个特征来区分良性模型和后门模型。它们分别揭示了三个特点: 第一, 被攻击的目标类所对应的热力图突出了触发器的位置, 其具有更小的稀疏性; 第二, 触发器模式通常是集中的, 其不会分散到不成组的像素中; 第三, 与目标标签相关的热力图在不同的输入数据上会重复出现。作者对这三个指标进行综合考虑, 使用离群点检测算法确定目标模型的热力图是否存在异常。

DL/DF-TND。Wang 等人[50]讨论了数据有限和无数据情况下的后门检测, 提出了 DL-TND (Data-Limited TrojanNet Detector)和 DF-TND (Data-Free TrojanNet Detector)来应对这两种情况。作者通过探索后门攻击与对抗样本攻击的关联, 并使用特征反演 (Feature Inversion)等技术, 设计了一套后门检测方法, 可以在只使用少量干净数据 (每个类一张图

片)和无数据的情况完成后门检测。

4.2 小结

模型级防御涵盖了多种防御方法, 如逆向触发器、元分类器和模型修复等。基于逆向触发器的方法通过逆向攻击中使用的触发器发现攻击, 并利用生成的触发器信息辅助后门防御。触发器信息为防御者提供了后门知识, 使后续防御更加精准和高效。基于元分类器的方法利用木马集训练后门模型, 根据不同的特征提取技术构建可靠的元分类器。这类方法的最大特点是检测器运行时效率高, 在实际应用中更加实用, 但其前期准备成本相对较高, 且对木马集的依赖性较大, 所以在面对强攻击时防御能力显出不足。模型修复类方法专注后门的移除工作。简单的修复方案往往会导致模型在干净数据上的准确性显著下降, 虽然先进的方法在模型性能和防御效果之间实现了很好的权衡, 但仍无法应对较强的攻击。此外, 这类方法虽然有众多研究工作, 但目前仍没有一个较为统一的框架可以集成发挥各工作的优势以形成一套完整的后门检测及修复两阶段防御体系。

5 输入级防御方法

输入级防御和模型级防御一样应用于模型训练后, 其也考虑防御方从第三方获取模型的场景。不过, 输入级防御的关注点主要落在神经网络模型在推理时的异常输入, 通过分析触发输入和干净输入在后门模型上的不同性质和表现, 从检测触发输入、修复触发输入、输入预处理等方面防御后门攻击。表 3 按时间顺序总结对比了主要的输入级防御方法的原理及防御手段。

表 3 输入级防御方法

Table 3 Defense at input level

方法	时间	原理/简述	防御手段	防御场景	防御效果
自编码器预处理[51]	2017	在输入数据和目标模型之间放置自编码器	输入预处理	从第三方获取模型	对于简单的木马攻击, 可使 90.2%的触发器失效
STRIP 检测系统[54]	2019	对输入图像叠加图像模式	检测触发输入	从第三方获取模型	可部署在黑盒场景下, 有效检测常见像素攻击和图像混合攻击
深度概率模型[55]	2019	创建深度概率模型, 通过相关指标识别触发输入	检测触发输入	从第三方获取模型	可有效检测普通像素攻击
SentiNet 检测框架[56]	2020	提取输入图像的高显著区域, 定位恶意触发器区域	检测触发输入	从第三方获取模型	可有效检测 BadNets、Trojan 等常见攻击, 以及物理上可实现的攻击
CLEANN 检测框架[57]	2020	通过字典学习和稀疏逼近揭示干净数据的统一行为	检测触发输入	从第三方获取模型	可在资源有限的嵌入式设备上有效检测 BadNets 和 Trojan 攻击
Februus 防御架构[58]	2020	识别恶意触发器区域, 利用 GAN 修复输入图像	修复触发输入	从第三方获取模型	对比神经清洁, 可检出更大尺寸、复杂和真实的触发器, 并有效修复触发输入
空间变换预处理[59]	2021	对输入图像进行翻转和缩放	输入预处理	从第三方获取模型	可有效检测带有静态触发器的攻击

5.1 方法介绍

一些工作关注如何在推理阶段直接检测潜在的触发输入。

STRIP 检测系统。Gao 等人[54]提出了 STRIP (STRong Intentional Perturbation)木马检测系统, 该方法的核心思想是对输入数据进行强扰动, 根据不同扰动下输入的预测变化来判断触发输入。在实验中, 作者尝试对输入图像叠加不同的图像模式。他们使用数据集中的其他图像对输入图像进行线性混合, 得到一组输入图像的副本。随后, 检查分类器在这些副本上的分类结果的随机性。由于鲁棒的后门触发器可以在图像被扰动后仍然能成功引导分类器将图像分类到目标标签, 而正常的输入则会因为一系列的强扰动无法被分类器识别, 所以低随机性往往指示着该输入图像是一个恶意的触发输入。

深度概率模型。Subedar 等人[55]提出, 可以为目标网络创建深度概率模型 (Deep Probabilistic Model)来量化输入数据的不确定性。其中, 深度概率模型揭示了 DNN 在干净数据上特征的概率分布。不确定性作为评价输入数据可疑程度的指标, 值越高代表对应的输入数据与概率模型越不相关, 进而表示当前输入有较大可能是恶意的触发输入。基于上述思想, 作者提出了两种方法。第一种方法使用干净的数据集学习网络深层特征的参数概率分布, 计算输入数据对这些深层特征分布的似然作为它们的不确定性估值。第二种方法利用平均场变分推断 (Mean-field Variational Inference)训练贝叶斯网络, 使用贝叶斯不一致主动学习 (Bayesian Active Learning by Disagreement)来测量不确定性。

SentiNet 检测框架。Chou 等人[56]提出 SentiNet

检测框架, 对物理上可实现的后门攻击进行了防御。在其研究的攻防场景下, 触发器是覆盖在图像上的一个小的连续区域。该方法将待预测样本输入目标网络, 根据触发器具有相邻区域强特征的性质, 利用可解释性技术和目标检测技术找到这些区域, 进而发现可能的触发输入。与前文所述的 STRIP[54]类似, 该方法认为触发输入通常是鲁棒的, 具有较强的抗扰动能力, 且触发器可以导致多个不同类别的输入分类错误, 所以, 可以考虑提取输入图像中高度显著的相连区域并将其覆盖在一组干净的图像上, 测试这些图像预测结果的变化情况来识别恶意的触发输入。

CLEANN 检测框架。Javaheripi 等人[57]利用字典学习和稀疏逼近等技术来揭示干净数据的统一行为, 从而鉴别恶意的触发输入。此外, CLEANN 是一种针对嵌入式应用的神经网络后门在线检测框架, 作者设计了专用的加速硬件来提高算法的执行效率。

在检测到触发输入之后, 还有一些研究提出方法可以将触发输入修复为干净输入。

Februus 防御架构。Februus[58]是 Doan 等人提出的一种即插即用的后门防御系统架构。在这种方法中, 作者采取移除恶意区域并重建图像的方法来修复触发输入, 进而防御后门攻击。在移除恶意区域阶段, 作者受到 SentiNet[55]检测方法的启发, 同样先寻找输入的高显著性区域再判定其是否为潜在触发器区域。Februus 采用了与 SentiNet 不同的方法判定触发器区域, 其使用一种自动选择灵敏度参数的方法完成判定。作者指出, 对于判定得到的触发器区域, 如果执行简单的删除操作, 会导致模型性能下降 10%左右。因此, Februus 通过“手术”重建这个区域, 首先使用中和色框替换触发器区域, 接着基于 GAN 进行图像修复, 重建被遮挡的图像区域。

此外, 还有研究工作提出, 可以通过对输入数据进行预处理, 来使触发输入失效。

自编码器预处理。早期, Liu 等人[51]提出使用自编码器作为防御后门攻击的手段。他们将自编码器放置在输入数据与模型之间, 作为输入数据的预处理器。由于经过干净数据所训练的自编码器能够记住正常数据的特征, 所以在推理阶段时, 干净数据在通过自编码器后会获得正常的数据表示, 因此 DNN 模型能够正常的对其进行预测, 而触发输入通过自编码器时其处理结果相对于原始输入会产生较大的偏差, 模型则难以识别触发输入, 从而使触发输入失效。

空间变换预处理。Li 等人[59]重新审视了主流的后门攻击方法, 发现大多数攻击采用了静态触发器的设置, 即在训练图像和测试图像上使用相同的、固定的触发器, 它们的外观和位置保持不变。于是, 作者考虑在测试图像通过目标网络之前, 采用翻转和缩放等图像空间变换方式对输入图像进行预处理。他们证明了这种方法能够一定程度上使得触发输入失效。同时, 该方法具有较高的效率, 且不需要外部数据或知识。然而, 作者也提出, 翻转和缩放等方式如果在训练阶段被用于中毒图像增强, 则基于这类预处理方法的防御手段可能会失效。

5.2 小结

输入级防御将后门攻击防御的重点放在模型推理阶段。在未知输入交给模型预测之前, 使用检测器或预处理器对数据进行评估或处理, 防止触发器进入模型激活后门。其中, SentiNet[56]和 Februus[58]方法验证了对物理可行攻击的防御。Februus 还在检测触发输入的基础上增加了输入修复方案, 使模型能够公平对待被篡改的数据。基于预处理的方法在不进行任何输入检测的情况下对所有输入进行统一变换, 破坏触发输入上的触发器信息, 并保证干净输入上信息的相对完整性。但目前研究尚不充分, 仅有少量相对简单的防御方案[51,59]。输入级防御属于被动防御, 其只能尽量过滤触发输入, 而不能有效检测模型是否真正被植入后门, 也不能对已植入后门的模型进行修复。一旦过滤失败, 后门依然会被触发。此外, 由于对所有测试样本都无差别的过滤或修复, 此方法还有一定的概率导致正常的测试样本被漏判或错判, 这在一些场景下可能是不可接受的。

6 可认证鲁棒性防御方法

可认证鲁棒性防御关注模型训练和部署的全生命周期, 旨在通过理论分析验证模型在某个攻击程度下具有严格的鲁棒性, 即模型在测试点周围具有统一的预测值。由于其鲁棒性验证是有严格理论保证的, 所以或可在一定条件下从根本上直接避免模型被植入后门。不过, 这类方法目前的研究尚处于起步阶段, 还有待进一步深入研究。表 4 按时间顺序总结了当前可认证鲁棒性防御的主要方法。

6.1 方法介绍

基于随机平滑的可认证鲁棒性。对抗样本攻击领域中的随机平滑 (Randomized Smoothing)[63]是可认证鲁棒性的典型研究。Wang 等人[60]首先对后门攻击中可认证鲁棒性防御进行了尝试。他们延用了对抗样本攻击中基于随机平滑的防御方法, 将其推广为对训练集和测试输入同时进行随机平滑, 从而创

表 4 可认证鲁棒性防御方法

Table 4 Defense based on robustness certification

分类	方法	时间	原理/简述	防御手段	防御场景	防御效果
随机平滑	文献[60]	2020	对训练集和测试输入进行随机平滑	提供模型保障	从第三方获取训练数据	在 MINIST 数据集的二分类任务中, 对图像进行任意 2 像素的扰动, 可保证有 36%的图像被正确分类
	RAB[61]	2023	一种更通用的验证模型鲁棒性的框架	提供模型保障	从第三方获取训练数据	在多种数据集的二分类任务中, 可对单像素、四像素和随机噪声攻击下的图像提供更高的认证精度
最近邻分类	文献[62]	2022	验证基于最近邻的分类器对于后门攻击的内在认证鲁棒性	提供模型保障	从第三方获取训练数据	可为普通像素攻击提供一定的鲁棒性保证

建平滑分类器。从实验上证明了利用随机平滑防御后门攻击是可行的, 但现有防御方法取得的效果比较有限, 仍需进一步研究。

与此同时, Weber 等人[61]也推广了随机平滑技术的应用, 提出了 RAB (Robustness Against Backdoor) 框架, 其可以验证针对后门攻击的模型鲁棒性。作者指出防御认证的目标是, 无论攻击者向训练集添加了什么样的触发器, 只要其扰动范围限制在半径为 R 的“L_p-Ball”范围内, 都能够确保模型的预测保持不变。与文献[60]相比, 他们基于不同的随机平滑噪声分布进行了更详细的评估、分析和实验, 并利用 Neyman-Pearson 引理严格推导了模型可认证的鲁棒性边界。该方法的具体实现过程如下。首先, 对于给定的训练集 $\boldsymbol{D}$, 根据一定的平滑分布添加随机噪声 $\boldsymbol{\epsilon_i}$ ($i \in \{1, ..., N\}$)生成 N 个平滑训练集 $\boldsymbol{D}+\boldsymbol{\epsilon_i}$; 其次, 使用 $\boldsymbol{D}+\boldsymbol{\epsilon_i}$ 得到 N 个平滑分类器; 最后, 在推理时, 再次对输入数据进行随机平滑生成多个副本数据, 并通过两次投票来得到最终模型的预测输出。

基于最近邻的可认证鲁棒性。最近, Jia 等人[62]提出可以利用最近邻 (Nearest Neighbor)算法为模型提供针对后门攻击的鲁棒性认证。他们推导了 kNN (K-Nearest Neighbors) 和 rNN (Radius Nearest Neighbors)针对后门攻击的内在认证鲁棒性。与先前基于随机平滑方法的防御不同, 基于最近邻的认证防御方法无须额外创建专门的投票机制, 而是利用算法自身固有的多数投票来预测数据标签, 进而提供鲁棒性认证。该方法不仅可以认证单个测试样本上的鲁棒性, 还可以对整个测试样本集合进行联合鲁棒性认证, 从而提高模型的整体认证精度。最近邻分类器用于图片分类时, 往往需要前期做特征提取来提取每张图片的特征向量, 如果特征提取器是基于神经网络的, 那么也可认为此研究与神经网络后门攻击有关。但本质上, 这项研究的关注的还是传统的基于近邻的分类方法的可认证鲁棒性研究, 由于神经网络本身具有较高的不可解释性, 所以这类研究中的鲁棒性保证是否能够比较容易地扩展到神经网络后门攻击中仍有待进一步研究。

此外, 还有一些针对数据投毒攻击的可认证鲁棒性研究[64-66], 由于数据投毒是实现后门攻击的主要手段之一, 所以这些研究对后门攻击的可认证鲁棒性防御也有一定的启发性。

6.2 小结

可认证鲁棒性防御从理论上保证了防御在一定条件下的成功, 因此不会被先进的自适应攻击所绕过, 具有很高的研究价值。但这类方法的研究难度相对较高, 目前研究集中在随机平滑、多数投票等方面, 可以应对的触发器模式仍然相对简单, 认证精度也有很大的提升空间。

7 现状与展望

目前, 后门攻击和防御的主要战场还落在计算机视觉领域, 但一方面计算机视觉领域中新的攻击手段不断涌现, 另一方面其他深度学习应用领域中的后门攻击也在逐年增多, 为此, 有必要有针对性地研究面向这类新方法、新领域攻击的防御手段, 同时也需要建立更通用、更易于推广的防御机制。此外, 目前后门防御方法思路各异、场景多样, 也需要有统一的评价基准来对比和评价各方法在具体任务中的适用性和有效性。

针对新型后门攻击的防御方法。数据集级、模型级、输入级和可认证鲁棒性防御四类防御方法分别适用于不同的防御场景和深度学习生命周期阶段, 防御者可以根据自身所具备的防御条件和所处的攻防环境选择合适的防御方法。不过, 目前的防御方法虽然一定程度上取得了还不错的防御效果, 但更为先进和隐蔽的攻击手段也层出不穷, 当前防御方法是否能够应对新型的攻击手段尚有待进一步验证。举例来说, 自监督学习作为特征提取器一定程度上可以防御后门攻击, 但最新研究表明, 自监督学习本身依然面临后门攻击的威胁[67]。如果现有防御方

法不能够有效防御新型攻击, 则仍需进一步研究有针对性的防御方法, 以提高防御效果。

其他领域后门防御方法。本文主要讨论了主流计算机视觉领域的后门防御方法。然而, 近年来神经网络后门攻击逐渐活跃于其他新领域和新场景, 包括自然语言处理[68-69]、声学信号处理[70]、联邦学习[71-72]、深度强化学习[73-74]和图神经网络[75]等。神经网络后门攻击对深度学习的各应用领域均构成了不同程度的安全威胁。但是, 目前针对这些新领域、新场景的后门防御研究尚不充分, 甚至仅处于起步阶段。因此, 有必要多关注其他领域的后门攻击和防御, 以未雨绸缪、加强防范, 建立更强的人工智能安全保障体系。

更通用的后门防御方法。近年来, 神经网络后门攻击发展迅速, 成果层出不穷。除了传统的静态、可见、无语义触发器, 也开始出现动态、不可见、有语义的触发器, 触发器的多样性给防御工作带来了很大的挑战。目前, 多数后门防御方法只关注于特定的触发器和有限的后门攻击手段, 对于不同类型的触发器和更先进的攻击手段很难实现有效防御。神经网络后门防御方法应增强通用性方面的研究, 以应对更多类型的触发器、更先进的攻击手段和更丰富的应用领域。基于元分类器的后门防御[38-40]一定程度上可以提供更高的通用性, 但此类方法目前研究工作不多, 其防御能力还有进一步提升的空间。此外, 可认证鲁棒性防御[60-62]可以从理论上提供鲁棒的安全保障, 但现有的研究工作取得的防御效果与传统的经验性防御方法相比还存在着很大的差距, 有必要进一步深入研究。

防御评价基准。目前, 神经网络后门防御在计算机视觉领域取得了一定成果, 但与防御者相比, 攻击者往往拥有对抗的主动权和更丰富的攻击手段, 因此, 防御者往往只能发挥出相对有限的水平。当前的多数防御方法主要面向特定的任务和场景进行设计, 一方面缺乏更多样的任务和场景上的理论和实验分析, 另一方面对方法的应用范围也没有具体明确的界定。此外, 在实验中, 由于对神经网络模型结构、后门攻击的相关配置和是否使用模型优化算法和数据增强等手段没有统一规定, 不同文献采用了不同的实验配置, 这也无利于不同防御方法之间的公平比较。为此, 有必要加强后门防御评价基准方面的研究, 以促进防御能力的进一步提升。

8 结论

深度学习的发展使得神经网络的应用不断增加, 而任何基于神经网络的应用都可能受到后门攻击的威胁, 后门防御的研究具有现实意义。本文首先阐述了神经网络后门攻击和防御的基础概念。然后, 围绕深度学习的全生命周期和神经网络后门攻击的发起情形, 将目前主流的神经网络后门防御方法分为数据集级、模型级、输入级和可认证鲁棒性防御四类, 进行了全面、详细的概括、分析和比较。最后, 基于神经网络后门防御的研究现状对该领域的未来方向做了预测展望。

参考文献

[1] Voulodimos A, Doulamis N, Doulamis A, et al. Deep Learning for Computer Vision: A Brief Review[J]. *Computational Intelligence and Neuroscience*, 2018, 2018: 7068349.

[2] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[3] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[C]. *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016: 173-182.

[4] Young T, Hazarika D, Poria S, et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. *IEEE Computational Intelligence Magazine*, 2018, 13(3): 55-75.

[5] Ji S L, Du T Y, Li J F, et al. Security and Privacy of Machine Learning Models: A Survey[J]. *Journal of Software*, 2021, 32(1): 41-67.
(纪守领, 杜天宇, 李进锋, 等. 机器学习模型安全与隐私研究综述[J]. *软件学报*, 2021, 32(1): 41-67.)

[6] Li M H, Jiang P P, Wang Q, et al. Adversarial Attacks and Defenses for Deep Learning Models[J]. *Journal of Computer Research and Development*, 2021, 58(5): 909-926.
(李明慧, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J]. *计算机研究与发展*, 2021, 58(5): 909-926.)

[7] Biggio B, Nelson B, Laskov P. Poisoning Attacks Against Support Vector Machines[C]. *The 29th International Conference on International Conference on Machine Learning*, 2012: 1467-1474.

[8] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C/OL]. *The 3rd International Conference on Learning Representations,* 2015. https://doi.org/10.48550/arXiv.1412.6572.

[9] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]. *The 25th USENIX Conference on Security Symposium*, 2016: 601-618.

[10] Gu T Y, Dolan-Gavitt B, Garg S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain[EB/OL]. 2017: arXiv: 1708.06733. http://arxiv.org/abs/1708.06733.

[11] Chen X Y, Liu C, Li B, et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning[EB/OL]. 2017: arXiv: 1712.05526. http://arxiv.org/abs/1712.05526.

[12] Liu Y F, Ma X J, Bailey J, et al. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks[C]. *Computer Vision -*

ECCV 2020: *16th European Conference*, *August 23-28*, *2020*, *Proceedings*, *Part X*, 2020: 182-199.

[13] Zou M H, Shi Y, Wang C L, et al. PoTrojan: Powerful Neural-Level Trojan Designs in Deep Learning Models[EB/OL]. 2018: arXiv: 1802.03043. http://arxiv.org/abs/1802.03043.

[14] Ji Y J, Zhang X Y, Ji S L, et al. Model-Reuse Attacks on Deep Learning Systems[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018: 349-363.

[15] Du W, Liu G S. A Survey of Backdoor Attack in Deep Learning[J]. *Journal of Cyber Security*, 2022, 7(3): 1-16.
(杜巍, 刘功申. 深度学习中的后门攻击综述[J]. *信息安全学报*, 2022, 7(3): 1-16.)

[16] Shokri R, Stronati M, Song C Z, et al. Membership Inference Attacks Against Machine Learning Models[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 3-18.

[17] Tran B, Li J, Mądry A. Spectral Signatures in Backdoor Attacks[C]. *The 32nd International Conference on Neural Information Processing Systems*, 2018: 8011-8021.

[18] Hayase J, Kong W H, Somani R, et al. SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics[EB/OL]. 2021: arXiv: 2104.11315. http://arxiv.org/abs/2104.11315.

[19] Chen B, Carvalho W, Baracaldo N, et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering[EB/OL]. 2018: arXiv: 1811.03728. http://arxiv.org/abs/1811. 03728.

[20] Chan A, Ong Y S. Poison as a Cure: Detecting & Neutralizing Variable-Sized Backdoor Attacks in Deep Neural Networks[EB/OL]. 2019: arXiv: 1911.08040. http://arxiv.org/abs/1911.08040.

[21] Hong S, Chandrasekaran V, Kaya Y, et al. On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping[EB/OL]. 2020: arXiv: 2002.11497. http://arxiv.org/abs/2002.11497.

[22] Borgnia E, Cherepanova V, Fowl L, et al. Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks without an Accuracy Tradeoff[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 3855-3859.

[23] Geiping J, Fowl L, Somepalli G, et al. What Doesn't Kill you Makes you Robust(Er): Adversarial Training Against Poisons and Backdoors[EB/OL]. 2021: arXiv: 2102.13624. http://arxiv.org/abs/2102.13624.

[24] Zeng Y, Park W, Mao Z M, et al. Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 16453-16461.

[25] Li Y G, Lyu X X, Koren N, et al. Anti-Backdoor Learning: Training Clean Models on Poisoned Data[EB/OL]. 2021: arXiv: 2110. 11571. http://arxiv.org/abs/2110.11571.

[26] Huang K Z, Li Y M, Wu B Y, et al. Backdoor Defense via Decoupling the Training Process[EB/OL]. 2022: arXiv: 2202.03423. http://arxiv.org/abs/2202.03423.

[27] Wang Y S, Ma X J, Chen Z Y, et al. Symmetric Cross Entropy for Robust Learning with Noisy Labels[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 322-330.

[28] Wang B L, Yao Y S, 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.

[29] Guo W B, Wang L, Xing X Y, et al. TABOR: A Highly Accurate Approach to Inspecting and Restoring Trojan Backdoors in AI Systems[EB/OL]. 2019: arXiv: 1908.01763. http://arxiv.org/abs/1908.01763.

[30] Liu Y Q, Lee W C, Tao G H, et al. ABS: Scanning Neural Networks for Back-Doors by Artificial Brain Stimulation[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 1265-1282.

[31] Shen G Y, Liu Y Q, Tao G H, et al. Backdoor Scanning for Deep Neural Networks through K-Arm Optimization[EB/OL]. 2021: arXiv: 2102.05123. http://arxiv.org/abs/2102.05123.

[32] Zhu L W, Ning R, Wang C, et al. GangSweep: Sweep out Neural Backdoors by GAN[C]. *The 28th ACM International Conference on Multimedia*, 2020: 3173-3181.

[33] Chen H L, Fu C, Zhao J S, et al. DeepInspect: A Black-Box Trojan Detection and Mitigation Framework for Deep Neural Networks[C]. *The Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019: 4658-4664.

[34] Veldanda A K, Liu K, Tan B, et al. NNoculation: Catching BadNets in the Wild[EB/OL]. 2020: arXiv: 2002.08313. http://arxiv.org/abs/2002.08313.

[35] Qiao X M, Yang Y K, Li H. Defending Neural Backdoors via Generative Distribution Modeling[EB/OL]. 2019: arXiv: 1910.04749. http://arxiv.org/abs/1910.04749.

[36] Tao G H, Shen G Y, Liu Y Q, et al. Better Trigger Inversion Optimization in Backdoor Scanning[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13358-13368.

[37] Dong Y P, Yang X, Deng Z J, et al. Black-Box Detection of Backdoor Attacks with Limited Information and Data[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 16462-16471.

[38] Kolouri S, Saha A, Pirsiavash H, et al. Universal Litmus Patterns: Revealing Backdoor Attacks in CNNS[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 298-307.

[39] Huang S, Peng W Q, Jia Z W, et al. One-Pixel Signature: Characterizing CNN Models for Backdoor Detection[C]. *Computer Vision - ECCV 2020*: *16th European Conference*, *Part XXVII*, 2020: 326-341.

[40] Xu X J, Wang Q, Li H C, et al. Detecting AI Trojans Using Meta Neural Analysis[C]. *2021 IEEE Symposium on Security and Privacy*, 2021: 103-120.

[41] Liu K, Gavitt B D, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks[C]. *The International Symposium on Research in Attacks, Intrusions, and Defenses*, 2018: 273-294.

[42] Zhao P, Chen P Y, Das P, et al. Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness[EB/OL]. 2020: arXiv: 2005.00060. http://arxiv.org/abs/2005.00060.

[43] Li Y G, Lyu X X, Koren N, et al. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks[EB/OL]. 2021: arXiv: 2101.05930. http://arxiv.org/abs/2101.05930.

[44] Jiang W, Wen X Y, Zhan J Y, et al. Interpretability-Guided Defense Against Backdoor Attacks to Deep Neural Networks[J]. *IEEE*

Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41(8): 2611-2624.

[45] Wu D X, Wang Y S. Adversarial Neuron Pruning Purifies Backdoored Deep Models[EB/OL]. 2021: arXiv: 2110.14430. http://arxiv.org/abs/2110.14430.

[46] Zhao Y, Zhu H, Chen K, et al. AI-Lancet: Locating Error-Inducing Neurons to Optimize Neural Networks[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 141-158.

[47] Sun B, Sun J, Pham L H, et al. Causality-Based Neural Network Repair[C]. *2022 IEEE/ACM 44th International Conference on Software Engineering*, 2022: 338-349.

[48] Zeng Y, Chen S, Park W, et al. Adversarial Unlearning of Backdoors via Implicit Hypergradient[EB/OL]. 2021: arXiv: 2110.03735. http://arxiv.org/abs/2110.03735.

[49] Huang X J, Alzantot M, Srivastava M. NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations[EB/OL]. 2019: arXiv: 1911.07399. http://arxiv.org/abs/1911.07399.

[50] Wang R, Zhang G Y, Liu S J, et al. Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free Cases[C]. *Computer Vision - ECCV 2020*: *16th European Conference*, *Part XXIII*, 2020: 222-238.

[51] Liu Y T, Xie Y, Srivastava A. Neural Trojans[C]. *2017 IEEE International Conference on Computer Design*, 2017: 45-48.

[52] Garipov T, Izmailov P, Podoprikhin D, et al. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNS[C]. *The 32nd International Conference on Neural Information Processing Systems*, 2018: 8803-8812.

[53] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[EB/OL]. 2015: arXiv: 1503.02531. http://arxiv.org/abs/1503.02531.

[54] Gao Y S, Xu C G, Wang D R, et al. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks[C]. *The 35th Annual Computer Security Applications Conference*, 2019: 113-125.

[55] Subedar M, Ahuja N, Krishnan R, et al. Deep Probabilistic Models to Detect Data Poisoning Attacks[EB/OL]. 2019: arXiv: 1912.01206. http://arxiv.org/abs/1912.01206.

[56] Chou E, Tramèr F, Pellegrino G. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems[C]. *2020 IEEE Security and Privacy Workshops*, 2020: 48-54.

[57] Javaheripi M, Samragh M, Fields G, et al. CleaNN: Accelerated Trojan Shield for Embedded Neural Networks[C]. *The 39th International Conference on Computer-Aided Design*, 2020: 1-9.

[58] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems[C]. *The 36th Annual Computer Security Applications Conference*, 2020: 897-912.

[59] Li Y M, Zhai T Q, Wu B Y, et al. Rethinking the Trigger of Backdoor Attack[EB/OL]. 2020: arXiv: 2004.04692. http://arxiv.org/abs/2004.04692.

[60] Wang B H, Cao X Y, jia J Y, et al. On Certifying Robustness Against Backdoor Attacks via Randomized Smoothing[EB/OL]. 2020: arXiv: 2002.11750. http://arxiv.org/abs/2002.11750.

[61] Weber M, Xu X J, Karlaš B, et al. RAB: Provable Robustness Against Backdoor Attacks[C]. *2023 IEEE Symposium on Security and Privacy*, 2023: 1311-1328.

[62] Jia J Y, Liu Y P, Cao X Y, et al. Certified Robustness of Nearest Neighbors Against Data Poisoning and Backdoor Attacks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(9): 9575-9583.

[63] Cohen J M, Rosenfeld E, Kolter J Z. Certified Adversarial Robustness via Randomized Smoothing[EB/OL]. 2019: arXiv: 1902.02918. http://arxiv.org/abs/1902.02918.

[64] Rosenfeld E, Winston E, Ravikumar P, et al. Certified Robustness to Label-Flipping Attacks via Randomized Smoothing[C]. *The 37th International Conference on Machine Learning*, 2020: 8230-8241.

[65] Levine A, Feizi S. Deep partition aggregation: Provable defenses against general poisoning attacks[C/OL]. *The 9th International Conference on Learning Representations*, 2021. https://doi.org/10.48550/arXiv.2006.14768.

[66] Jia J Y, Cao X Y, Gong N Z. Intrinsic Certified Robustness of Bagging Against Data Poisoning Attacks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(9): 7961-7969.

[67] Jia J Y, Liu Y P, Gong N Z. BadEncoder: Backdoor Attacks to Pre-Trained Encoders in Self-Supervised Learning[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 2043-2059.

[68] Chen X Y, Salem A, Chen D F, et al. BadNL: Backdoor Attacks Against NLP Models with Semantic-Preserving Improvements[C]. *The 37th Annual Computer Security Applications Conference*, 2021: 554-569.

[69] Chen K J, Meng Y X, Sun X F, et al. BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models[C/OL]. *The 10th International Conference on Learning Representations*, 2022. https://doi.org/10.48550/arXiv.2110.02467.

[70] Zhai T Q, Li Y M, Zhang Z Q, et al. Backdoor Attack Against Speaker Verification[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 2560-2564.

[71] Xie C L, Huang K L, Chen P Y, et al. DBA: Distributed backdoor attacks against federated learning[C/OL]. *The 8th International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=rkgyS0VFvr.

[72] Wang H Y, Sreenivasan K, Rajput S, et al. Attack of the Tails: Yes, you Really can Backdoor Federated Learning[EB/OL]. 2020: arXiv: 2007.05084. http://arxiv.org/abs/2007.05084.

[73] Wang Y, Sarkar E, Li W Q, et al. Stop-and-Go: Exploring Backdoor Attacks on Deep Reinforcement Learning-Based Traffic Congestion Control Systems[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 4772-4787.

[74] Wang L, Javed Z, Wu X, et al. BACKDOORL: Backdoor Attack Against Competitive Reinforcement Learning[EB/OL]. 2021: arXiv: 2105.00579. http://arxiv.org/abs/2105.00579.

[75] Xi Z H, Pang R, Ji S L, et al. Graph backdoor[C]. *The 30th USENIX Security Symposium*, 2021: 1523-1540.

江钦辉 于 2020 年在怀化学院计算机科学与技术专业获得学士学位。2023 年于广州大学“方班”获硕士学位。研究领域为人工智能安全。研究兴趣包括: 人工智能安全。Email: qinhui.jiang@outlook.com

孙彦斌 于 2016 年在哈尔滨工业大学计算机科学与技术专业获得博士学位。CCF 高级会员。现任广州大学网络空间安全学院教授。研究兴趣包括: 网络安全、工控安全、未来网络。Email: sunyanbin@gzhu.edu.cn

李默涵 于 2016 年在哈尔滨工业大学计算机软件与理论专业获得博士学位。CCF 高级会员。现任广州大学网络空间安全学院教授。研究兴趣包括: 人工智能安全、数据质量、入侵检测。Email: limohan@gzhu. edu.cn