计算机科学
COMPUTER SCIENCE

一种基于深度分区聚合的神经网络后门样本过滤方法

郭嘉铭, 杜文韬, 杨超

引用本文

郭嘉铭, 杜文韬, 杨超. 一种基于深度分区聚合的神经网络后门样本过滤方法[J]. 计算机科学, 2025, 52(11): 425-433.

GUO Jiaming, DU Wentao, YANG Chao. Neural Network Backdoor Sample Filtering Method Based on Deep Partition Aggregation [J]. Computer Science, 2025, 52(11): 425-433.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900007

一种基于深度分区聚合的神经网络后门样本过滤方法

郭嘉铭[1]　杜文韬[1]　杨　超[2,3]
1 湖北大学网络空间安全学院　武汉 430062
2 湖北大学计算机学院　武汉 430062
3 智慧政务与人工智能应用湖北省工程研究中心　武汉 430062
(1196951311@qq.com)

摘　要　深度神经网络易受后门攻击,攻击者可以通过数据投毒的方式植入后门并劫持模型的行为。其中,类特定攻击映射关系复杂、与正常任务关联紧密,因而能绕过大多数防御方法,具有更高的威胁性。文中研究了类特定攻击在植入后门的过程中攻击成功率与模型分类性能的关系,总结出3条性质,并以此为基础设计了一种针对类特定攻击的样本过滤方法。该方法使用深度分区聚合(Deep Partition Aggregation,DPA)的集成学习方法与投票法对数据集进行反复迭代过滤。根据类特定攻击的3条性质,从数学层面证明了该过滤方法的有效性,并在标准分类数据集上进行了大量实验,在迭代4轮后均能过滤95%以上的后门样本。同时,与最新的样本过滤方法的对比实验结果,体现了所提过滤方法在针对类特定攻击时的优越性。文中实验基于Github的开源项目backdoorbox开展。

关键词:深度学习;数据投毒;后门攻击;类特定攻击;集成学习;样本过滤
中图分类号　TN915.08

Neural Network Backdoor Sample Filtering Method Based on Deep Partition Aggregation

GUO Jiaming[1],DU Wentao[1] and YANG Chao[2,3]
1 School of Cyber Science and Technology,Hubei University,Wuhan 430062,China
2 School of Computer Science,Hubei University,Wuhan 430062,China
3 Engineering Research Center of Hubei Province in Intelligent Government Affairs and Application of Artificial Intelligence,Wuhan 430062,China

Abstract　Deep neural networks are vulnerable to backdoor attacks,where attackers can implant backdoors and hijack model behavior by poisoning data. Among them,class-specific attacks can bypass most defense methods due to their complex mapping relationships and close association with normal tasks,making them more threatening. This paper studies the relationship between attack success rate and model classification performance in the process of implanting backdoors for class-specific attacks,summarizes three properties,and designs a sample filtering method based on these properties to address class-specific attacks. This method uses the Deep Partition Aggregation(DPA) ensemble learning method and voting method to iteratively filter backdoor samples. This paper mathematically proves the effectiveness of this filtering method based on three properties of class-specific attacks,and conducts extensive experiments on standard classification datasets. After four iterations,it filters more than 95% of backdoor samples in all experiments. At the same time,the results of comparative experiments with the latest sample filtering methods,demonstrate the superiority of proposed method in addressing class-specific attacks. The experiments in this paper are based on the open-source project backdoorbox on Github.

Keywords　Deep learning,Data poisoning,Backdoor attack,Class-specific attacks,Ensemble learning,Sample filtering

1　引言

随着人工智能的日益普及,对高级智能的需求导致模型开发过程愈发复杂。训练一个强大的深度学习模型通常需要大量的训练数据和计算资源,这对于普通用户或小企业而言都是不小的负担。因此,利用开源数据集[1]、开源预训练模型[2]、第三方模型训练平台[3]等第三方资源缓解资源不足的问题是十分常见的做法,但这种行为同时存在引入恶意第三方实体的可能,会导致深度学习攻击面的增加[4-5]。

目前深度学习领域面临多种安全威胁,其中后门攻击

到稿日期:2024-09-02　返修日期:2024-11-21
基金项目:国家自然科学基金(61977021);湖北省重点研发计划(2021BAA188)
This work was supported by the National Natural Science Foundation of China(61977021) and Key R&D Program of Hubei Province,China(2021BAA188).
通信作者:杨超(stevenyc@hubu.edu.cn)

(Backdoor Attack)作为一种在机器学习管道中全阶段可用且效果斐然的攻击方法[6],对深度学习领域具有不容忽视的威胁性,如图1所示。在推理过程中,攻击者可以通过在输入数据中插入对应后门的触发器来劫持模型预测[7]。对防御者而言,由于后门与触发器都具备隐蔽的特征[5],难以分辨被植入后门的模型与带有触发器的数据,因此防御后门攻击绝非易事。

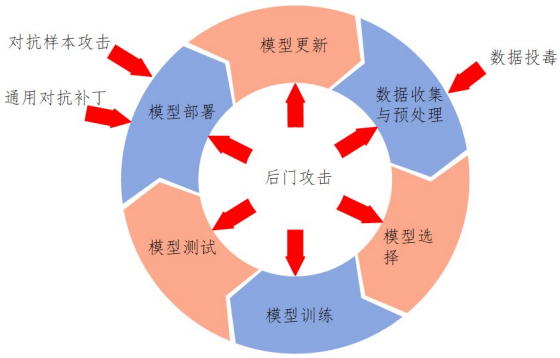

图1 机器学习管道与可能遭受的攻击

Fig. 1 Machine learning pipelines and potential attacks

根据攻击阶段的不同,后门攻击也有许多种类[6],其中基于数据投毒的后门攻击是最简单的在训练过程中完成后门植入的方法[5],也是目前实现后门攻击最常用的方法[8]。根据攻击是否依赖样本源类,可将基于数据投毒的后门攻击划分为"类特定(Class-specific)"攻击与"类不可知(Class-agnostic)"攻击[6](也称为"All-to-All Attack"与"Singal Target Attack"[9])。其中,类特定攻击因其映射关系复杂、后门与正常任务关联性强的特点,可以绕过 Neural Cleanse[10]、B3D[11]、SentiNet[12]等多种经典后门防御方法,最新的 SCALE-UP[13]、IBD-PSC[14]等过滤方法也收效甚微,因此具有更大的威胁性[5]。本文针对类特定攻击进行研究分析,并提出一种样本过滤方法,具体工作如下。

1)理论分析。本文认为类特定攻击以"识别样本源类"为必要条件的约束必然导致训练过程中攻击成功率(Attack Success Rate,ASR)对干净数据分类准确率(Clean Data Accuracy,CDA)的强依赖性。此外,由于后门植入基本不影响模型的正常分类性能,因此在模型性能良好的前提下,CDA应当大于ASR。实验数据也支持这一结论。同时,由于干净数据集通常存在较高程度的冗余,即使对训练数据进行一定程度的分割,依旧可以在CDA下降不明显的前提下观察到这一现象。基于这一现象,本文总结出类特定攻击的3条性质,并证明可以通过样本过滤的方法实现对类特定攻击的防御。

2)防御方法。基于过往研究提出了一种针对类特定攻击的样本过滤方法。该方法通过 DPA[15]与投票法放大 CDA 与 ASR 的差值,可以通过不断迭代,在黑盒环境下分离出数据集中的绝大部分后门样本。本文在 CIFAR10[16]与 GTSRB[17]等两个标准分类数据集上分别使用 BadNets[9]、Blended[18]、PhysicalBA[19]、WaNet[20]这4种经典投毒方法,实现了循环移位的类特定攻击,并使用所提防御方法进行样本过滤,最终在迭代4轮后均达到了95%以上的后门样本过滤率。这一实验结果证明了在防御者能控制训练数据集的情况下,通过数据投毒进行的后门攻击难以使用同一触发器实现源类到目标类的复杂映射关系。

2 相关工作

2.1 基于数据投毒的后门攻击

2017年,Gu 等提出了 BadNets[9],通过标记一部分训练样本并修改其标签的方式,在训练过程中为模型植入后门。学术界由此开始广泛关注深度学习后门攻防的问题。Chen 等提出了 Blended[18],系统地测试了不同的后门攻击策略;Li 等提出了 PhysicalBA[19],研究了数字图像触发器在物理世界中的脆弱性并讨论了缓解方法;Nguyen 等提出了 WaNet[20],该攻击方法基于图像扭曲生成人眼难以分辨的触发器,完成了不可察觉的后门攻击;Doan 等提出的 LIRA[21]则在此基础上将触发器改进得更为隐蔽;Souri 等提出的 Sleeper Agent[22]则使用了一种梯度对齐的方法来进行后门攻击。总体而言,投毒后门攻击的发展呈现触发器隐蔽化、动态化、抗干扰的趋势。

2.2 基于样本过滤的后门防御方法

机器学习管道的不同阶段有不同的后门防御方法[6]。在数据收集阶段以样本过滤为主,这类方法同样也被应用于防御一般的投毒攻击[8]。基于样本过滤的后门防御方法通常从特征统计、激活情况、触发器鲁棒性等方面入手。

在特征统计方面,Tran 等[23]首先讨论了如何从数据集中过滤恶意样本,证明了中毒样本倾向于在特征表示的协方差谱中留下可检测的痕迹;Hayase 等提出了 SPECTRE[24],利用鲁棒协方差估计增强损坏数据的频谱特征;Javaheripi 等提出了 CLEANN[7],利用字典学习和稀疏逼近来表征良性数据的统计行为并识别木马触发器;Zeng 等[25]对图像频域进行了分析,发现许多后门攻击表现出严重的高频伪影,并且这些伪影在不同的数据集和分辨率中持续存在;Huang 等[26]通过认知蒸馏从图像中提取可导致相同模型输出的小图像(称为"认知模式"),发现后门样本的认知模式都异常的小。

在激活方面,Amarnath 等提出了 TESDA[27],利用干净样本与后门样本在深度神经网络中间层特征分布中引起的差异来在线检测攻击;Chen 等提出了激活聚类[28]的防御方法,该方法通过分析训练数据的神经网络激活情况对训练样本进行聚类,根据聚类结果判断训练数据是否中毒;Hou 等根据后门样本与良性样本在调整模型部分参数时预测一致性上的差异,设计了 IBD-PSC[14]。

在触发器鲁棒性方面,Liu 等提出了 TROJDEF[29],认为当向样本加入噪声时,直觉上后门输入更加稳定,因此通过监测预测置信度的变化来识别和过滤特洛伊木马的输入;Chen 等[30]观察到带有触发器的有毒样本的特征表示比干净样本的特征表示对转换更敏感,基于此设计了称为"特征一致性转换(FCT)"的灵敏度度量,以区分后门样本与干净样本;Guo 等提出了 SCALE-UP[13],该方法基于一种现象,即当放大所有像素值时,有毒样本的预测明显比良性样本的预测更加一致。

基于样本过滤的后门防御方法普遍是经验防御,缺乏理论依据,且类特定攻击与类不可知攻击在诸多特征上大相径庭,因此普遍存在对类特定攻击防御的空白;具备理论基础的可验证防御则对防御者能力或防御条件等提出了相对更严苛的要求,难以在黑盒环境下运行。

3 防御策略

表1列出了文中用到的符号及其含义。

表1 符号及含义

Table 1 Symbols and their meanings

符号	含义
D	训练集
$\boldsymbol{D}_c$	训练集中干净样本的集合
$\boldsymbol{D}_p$	训练集中后门样本的集合
$\lvert D\rvert$	训练集大小
p	投毒率，$p=\frac{\lvert D_p\rvert}{\lvert D\rvert}$
CDA	模型对干净数据的分类准确率
ASR	模型对后门样本的分类准确率

3.1 类特定攻击的性质分析

基于直觉与实验数据，本文总结出类特定攻击的3条性质。

1)对性能良好且已稳定的带有类特定攻击后门的模型而言，有：

$$CDA>ASR \tag{1}$$

这一性质来源于一种猜想：模型原本的分类任务只需识别出样本的源类，而类特定攻击必须在识别出源类的基础上识别出触发器，并正确触发后门，将样本的标签由源类映射到攻击者预期的目标类。因此，在忽略误报率的前提下，可以认为：

$$P(\text{类特定攻击成功})=P(\text{识别源类})\cdot P(\text{识别触发器}\mid\text{识别源类})\cdot P(\text{触发后门}\mid\text{识别触发器且识别源类}) \tag{2}$$

其中，类特定成功概率即模型的ASR，识别源类的概率即CDA。因此，模型的ASR应始终小于CDA，并且可以推测在训练过程相同时，后门攻击的触发器越隐蔽，源类到目标类的映射关系越复杂，ASR应当越低。

本文使用BadNets在MNIST[31]分类任务上植入循环移位后门。一个攻击者理想的被植入循环移位后门的深度学习分类器F对源标签为l的良性输入s的分类结果应当满足：

$$F(s)=l \tag{3}$$

$$F(s\oplus t)=(l+1)\bmod L \tag{4}$$

其中，t为该循环移位后门的触发器，L为F的输出层节点数。

选择基于BadNets实现循环移位后门的原因是BadNets所使用的触发器易于识别，循环移位逻辑也较为简单，因此可以尽量降低植入与触发后门的难度，最大程度地提高后门模型的ASR。若类特定后门具备上文所述性质，则在实验中，即使对如此简易的后门，也应当观察到ASR总是小于CDA的现象。

实验结果验证了这一猜想，如图2所示。

由于分类任务与后门逻辑均很简单，因此模型的CDA与ASR在训练中很早便达到稳定。但在20次采样中，ASR均小于CDA。这与BadNets实现类不可知攻击时的情况形成鲜明的对比，如图3所示。

可以看出，在基于BadNets实现类不可知攻击时，情况正好相反。在其他条件均相同的情况下，类不可知攻击的ASR反而总是大于CDA。这一对比证明了验证实验的有效性。

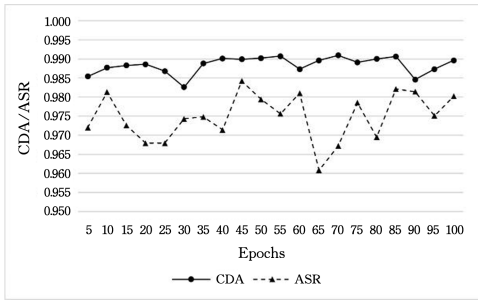

图2 BadNets在MNIST上植入循环移位后门的CDA与ASR

Fig. 2 CDA and ASR when BadNets implants a cyclic shift backdoor on MNIST

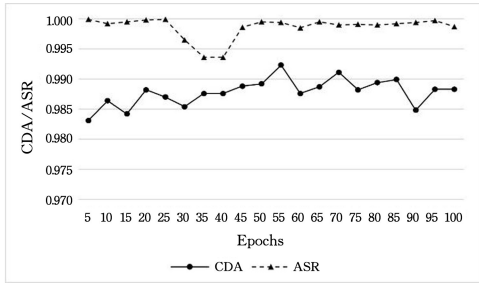

图3 BadNets在MNIST上实现类不可知攻击的CDA与ASR

Fig. 3 CDA and ASR when BadNets implants a class-agnostic backdoor on MNIST

2)在模型与其他参数均相同时，若两个模型的训练集满足：

$$p_1<p_2 \tag{5}$$

则有：

$$CDA_1=CDA_2,ASR_1<ASR_2 \tag{6}$$

即实现后门攻击的过程中，投毒率应只影响后门攻击的成功率，而不应影响模型对干净样本的分类准确率。

这一性质源于后门攻击本身固有的特点，即成功的后门植入不应影响分类模型对干净样本的分类能力。这也是基于数据投毒的后门攻击与通常意义上的数据投毒的差别所在。否则即使使用者不知道后门攻击的存在，模型性能达不到预期也会使其放弃该模型，后门植入变得毫无意义。因此，成功的投毒后门攻击不应使模型的CDA降低。本文分别测试了BadNets以0.05，0.02，0.01，0.005的投毒率对MNIST进行攻击的情况，结果如图4所示。

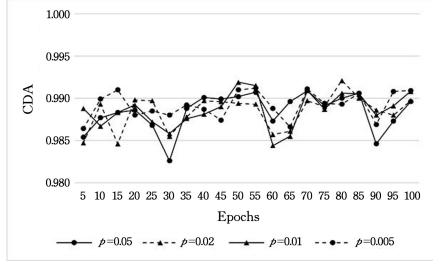

图4 BadNets在MNIST上以不同投毒率植入循环移位后门的CDA

Fig. 4 CDA when BadNets implants a cyclic shift backdoor at different poisoning rates on MNIST

不难看出，模型的CDA并没有因投毒率的降低而出现明显差别。与之相对的是模型的ASR，实验结果也符合投毒率降低引起攻击成功率降低的直觉，如图5所示。随着投毒率的降

低，模型在相同 epoch 上的攻击成功率出现了明显的差异。

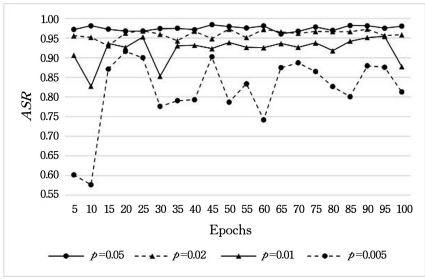

图 5 BadNets 在 MNIST 上以不同投毒率植入循环移位后门的 ASR

Fig. 5 ASR when BadNets implants a cyclic shift backdoor at different poisoning rates on MNIST

3)在模型与其他参数均相同时，存在$\varepsilon_c<\varepsilon_p<1$，若两个模型的训练集满足：

$$\varepsilon_p>\frac{|D_1|}{|D2|}>\varepsilon_c \tag{7}$$

则有：

$$\frac{CDA_2}{CDA_1}>\frac{ASR_2}{ASR_1} \tag{8}$$

即干净样本相较于后门样本具有更大的冗余性，因此当训练数据量足以使模型完成正常的训练时，模型在正常任务上具备更强的鲁棒性，受到的影响将小于后门任务。

对于一个数据量充足的训练集，在一定范围内减少其数据量并不会对模型的正常任务与后门植入产生实质性影响，此处假设对模型产生实质性影响的训练数据量比例阈值为ε_p；若训练数据量过少，则模型的 CDA 与 ASR 无法达到稳定，不具备研究意义，此处假设模型无法收敛的训练数据量比例阈值为ε_c。在以上述阈值范围内的数据量进行训练的情况下，模型在正常任务与后门植入上均能达到稳定，同时其 CDA 与 ASR 相比使用全部数据进行训练的情况均有一定程度的下降。假设使用全部数据D_1训练而得的性能参数为CDA_1与ASR_1，使用数据量在阈值范围内的子集D_2而得的性能参数为CDA_2与ASR_2，则有：

$$CDA_2<CDA_1,ASR_2<ASR_1 \tag{9}$$

与此同时，由于与性质 1 的原理相同，即后门的鲁棒性也以正常任务的鲁棒性为基础，因此 CDA 的下降幅度应低于 ASR，即：

$$\frac{CDA_1-CDA_2}{CDA_1}<\frac{ASR_1-ASR_2}{ASR_1} \tag{10}$$

故而得此性质。

原始数据集的冗余性是攻击者通过数据投毒实现后门攻击的基础，也是防御者通过样本过滤防御后门攻击的基础：攻击者可以将一部分干净样本修改为后门样本，而不必担心影响模型性能；防御者也可以在过滤后门样本时接受干净样本一定程度的损失。对本文过滤方法而言，可以避免在迭代中出现干净样本减少与基分类器 CDA 降低的恶性循环，相较于后门样本，干净样本更强的鲁棒性确保了 DPA 与投票法的有效性。

本文分别以 MNIST 数据集的全部数据、1/3 数据、1/10 数据与 1/20 数据使用 BadNets 实现循环移位后门，实验结果如图 6、图 7 所示。

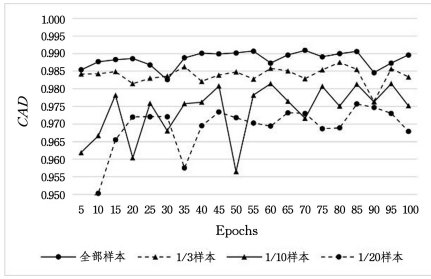

图 6 BadNets 在 MNIST 上使用不同比例的训练数据植入循环移位后门的 CDA

Fig. 6 CDA when BadNets implants a cyclic shift backdoor with different proportions of training data on MNIST

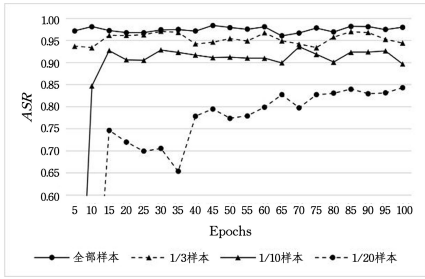

图 7 BadNets 在 MNIST 上使用不同比例的训练数据植入循环移位后门的 ASR

Fig. 7 ASR when BadNets implants a cyclic shift backdoor with different proportions of training data on MNIST

可以看到，随着训练样本的减少，CDA 受到的影响并不明显，ASR 则出现了明显的下降。在投毒率为 0.05、使用 1/20的数据量的情况下，CDA 相较于正常训练只降低了 2%左右，而 ASR 出现了约 15%的下滑。因此，在对 MNIST 做样本过滤时，损失少量数据对模型 CDA 造成的影响可以忽略不计。

3.2 基于 DPA 的后门样本过滤方法

本文方法的防御流程如图 8 所示。在后门样本过滤中，在迭代开始前设置两个集合：留集D_{in}与去集D_{out}，其中D_{in}为初始训练集 D，D_{out}为空集。算法的目标是通过多轮迭代使D_{out}中的后门样本尽可能多，干净样本则保持相对稳定的数量。在迭代过程中将通过 DPA 进行过滤，该方法分为分区训练与投票聚合两部分。

1)分区训练：将D_{in}分割为标签各不相同的集合，再将各子集均匀地分割为数个大小相等的子集，并以各子集为训练集分别训练基分类器。由于基分类器仅用于过滤后门样本，因此不需要一次训练所有标签，不同标签可以并行训练以提升效率。标签内的分割数量 k 应当满足性质 3)，即$\varepsilon_p>\frac{1}{k}>\varepsilon_c$。防御者无需知道具体的$\varepsilon_c$值，在黑盒条件下只需根据基分类器在实验中的表现选择一个合适的值，即首先确保基分类器的 CDA 不会使此轮过滤后所得的D_{in}数据量过少，导致后续迭代中无法获得稳定的基分类器。在此基础上，分割数量越多越好。

本文使用分区训练而非 Bagging[32] 的原因在于 Bagging 会产生大量重复参与基分类器训练的样本。鉴于分类器往往在其训练集上有更加良好的性能，后门样本重复参与训练极

有可能会增加其逃逸概率，因此选择分区训练的方法来避免这一现象。

2）投票聚合：通过投票法将分区训练所得的基分类器聚合为性能更强的大分类器。这些模型将对初始训练集 D 中的每个样本进行投票，根据投票结果决定该样本被划入 D_{in} 或 D_{out}。防御者可以自行制定完成 DPA 后的投票规则，例如，对任意样本，在 n 个基分类器中，若至少有 k 个基分类器的分类结果与其标签相同，则将该样本划入 D_{in}，否则划入 D_{out}。由于性质 1）与性质 3），对每个基分类器而言，均有 $CDA>ASR$，合适的投票机制会将 CDA 与 ASR 的比值成倍扩大。

投票结束后将判断迭代结束的条件是否被满足，若未结束则根据新划分的 D_{in} 进行下一轮迭代，若结束则将 D_{out} 中的所有数据视为后门样本并丢弃，保留 D_{in} 取代初始训练集 D。防御者可以根据具体需求调整迭代的停止条件，如 D_{in} 大小已经稳定，D_{out} 大小已达到阈值，迭代到达一定次数等。

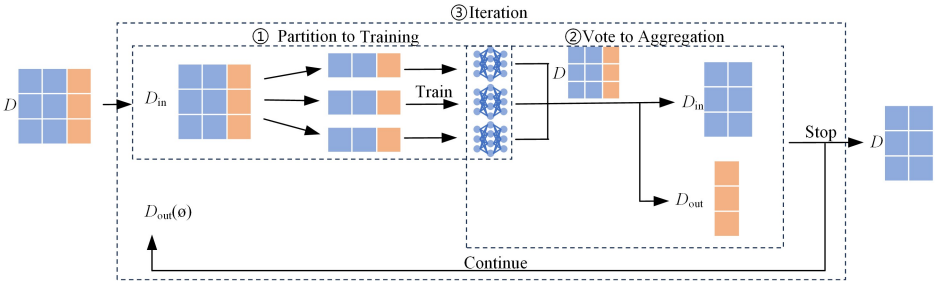

图 8 本文后门样本过滤方法的流程

Fig. 8 Flowchart of the proposed backdoor sample filtering method

3.3 有效性论证

令迭代过程中将 D_{in} 分割为 m 个子集进行训练，投票过程中任意样本至少有 k 个基分类器的预测结果与样本标签相同才会被划入 D_{in}。假设 DPA 中的基分类器具有相同的 CDA 与 ASR。

令第 n 轮迭代中，投票前 D_{in} 的投毒率为 p_n，投票后投毒率为 p_{n+1}，p_1 即训练集投毒率；投票前干净样本存活率为 C_n，后门样本存活率为 P_n，投票后分别为 C_{n+1} 与 P_{n+1}。则有：

$$C_{n+1}=\sum_{i=k}^{m}C_m^i\cdot CDA_n{}^i(1-CDA_n)^{m-i} \tag{11}$$

$$P_{n+1}=\sum_{i=k}^{m}C_m^i\cdot ASR_n{}^i(1-ASR_n)^{m-i} \tag{12}$$

由性质 1）可知，对任意基分类器均有 $CDA>ASR$，易知 $C_{n+1}>P_{n+1}$，从而可得：

$$p_{n+1}=p_1\cdot\frac{P_{n+1}}{p_1P_{n+1}+(1-p_1)C_{n+1}}<p_1 \tag{13}$$

即有 $p_2<p_1$。由性质 2）与性质 3）可知，$ASR_2<ASR_1$，$CDA_2=CDA_1$，从而可得：

$$C_3=C_2 \tag{14}$$

$$P_3<P_2 \tag{15}$$

$$p_3=p_1\cdot\frac{P_3}{p_1P_3+(1-p_1)C_3}<p_1\cdot\frac{P_2}{p_1P_2+(1-p_1)C_2}=p_2 \tag{16}$$

由此类推，可知对任意 n，满足：

$$p_{n+1}<p_n \tag{17}$$

即每轮迭代都将使 D_{in} 的投毒率降低，同时使干净样本存活率维持稳定值。

4 实验结果分析

4.1 实验设置

本文选取 BadNets，Blended，PhysicalBA，WaNet 这 4 种投毒后门攻击方法，在 CIFAR10 与 GTSRB 两个标准分类数据集上实现了对 ResNet-18 的循环移位类特定后门攻击，并使用本文方法对训练集做 4 轮迭代以进行过滤。其中，WaNet 的投毒率为 0.1，其余投毒方法的投毒率均为 0.05；迭代过程中将 CIFAR10 数据集划分为 5 份，将 GTSRB 数据集划分为 7 份；投票规则均为“若不多于两个基分类器分类错误则划入 D_{in}”。

4.2 后门检测

本文方法可以在一定程度上实现对类特定后门的检测。表 2 列出了数据集为 GTSRB、攻击为 BadNets 的情况下，第一轮迭代的基分类器训练结果对比。以 CDA 均值代入计算，经过第一轮迭代的投票后，两种情况下的 D_{out} 大小应分别占原训练集大小的约 0.15%与 0.48%，但在实际结果中，干净数据集下的 D_{out} 占原训练集大小的0.1502%，投毒数据集的 D_{out} 则占4.3243%，其中包括0.3942%的干净样本与91.31%的后门样本。基于如此显著的差异，防御者可以在黑盒情况下合理推断数据集受到了投毒后门攻击。

表 2 BadNets-GTSRB 干净训练集与投毒训练集的 DPA 结果

Table 2 Results of DPA on clean dataset and poisoned dataset of BadNets-GTSRB

(%)

	干净数据集	投毒数据集	
	CDA	*CDA*	*ASR*
M0	96.28	96.13	42.86
M1	96.29	93.55	33.00
M2	96.48	95.13	65.40
M3	97.14	94.00	40.30
M4	96.35	94.38	22.89
M5	95.95	92.63	27.79
M6	96.09	95.94	20.18
均值	96.37	94.54	36.06

4.3 过滤结果

过滤结果如表 3、表 4 所列，在迭代 2 轮的情况下均达到了 80%以上的后门样本过滤率，迭代 4 轮的情况下则均达到了 95%以上的过滤率。图 9 给出了分别使用过滤前后的训练集进行训练时，训练过程中的后门植入情况。在使用原本

的数据集能稳定植入后门的训练条件下，大部分完成后门过滤的训练集已无法完成后门的植入，少部分只能植入严重劣化的后门。

表 3 4 轮迭代后门样本的过滤率

Table 3 Filtration rate of backdoor samples in four iterations

(%)

Attack	Dataset	*Recall*_1	*Recall*_2	*Recall*_3	*Recall*_4
BadNets	CIRAR10	43.24	81.08	94.66	97.00
	GTSRB	83.44	96.34	96.34	96.34
Blended	CIRAR10	39.12	84.39	96.60	97.25
	GTSRB	77.30	98.46	98.46	98.46
PhysicalBA	CIRAR10	38.41	83.80	96.39	97.03
	GTSRB	49.25	86.56	95.52	95.52
WaNet	CIRAR10	76.87	94.53	95.93	96.78
	GTSRB	78.67	97.06	97.49	97.49

干净样本的损失是样本过滤中不可难免的问题。从表 4 中可以看出，虽然准确率较高，但精确率在部分组合下并不理想，在较为困难的 CIFAR10 分类任务上表现得最为明显。但基于干净数据的冗余性，一定程度的误报尚在可接受范围内。如图 10 所示，相较于使用原本训练集的模型，使用过滤后训练集的模型的 ASR 均有下降，但 CDA 并没有明显变化。实验结果证明了本文方法可以在不影响原训练任务的前提下过滤绝大部分后门样本。

表 4 第四轮迭代实验结果

Table 4 Experience result of the forth iteration

(%)

Attack	Dataset	*p*	*Recall*	*Precision*	*Accuracy*	*F*1
BadNets	CIRAR10	0.166	97.00	43.93	93.66	59.81
	GTSRB	0.181	96.34	96.76	99.67	96.55
Blended	CIRAR10	0.153	97.25	42.88	93.43	58.78
	GTSRB	0.081	98.46	58.63	96.45	73.50
PhysicalBA	CIRAR10	0.157	97.03	77.79	98.45	86.91
	GTSRB	0.223	95.52	93.32	99.77	94.41
WaNet	CIRAR10	0.392	96.78	52.76	90.88	66.76
	GTSRB	0.258	97.49	94.54	99.26	96.00

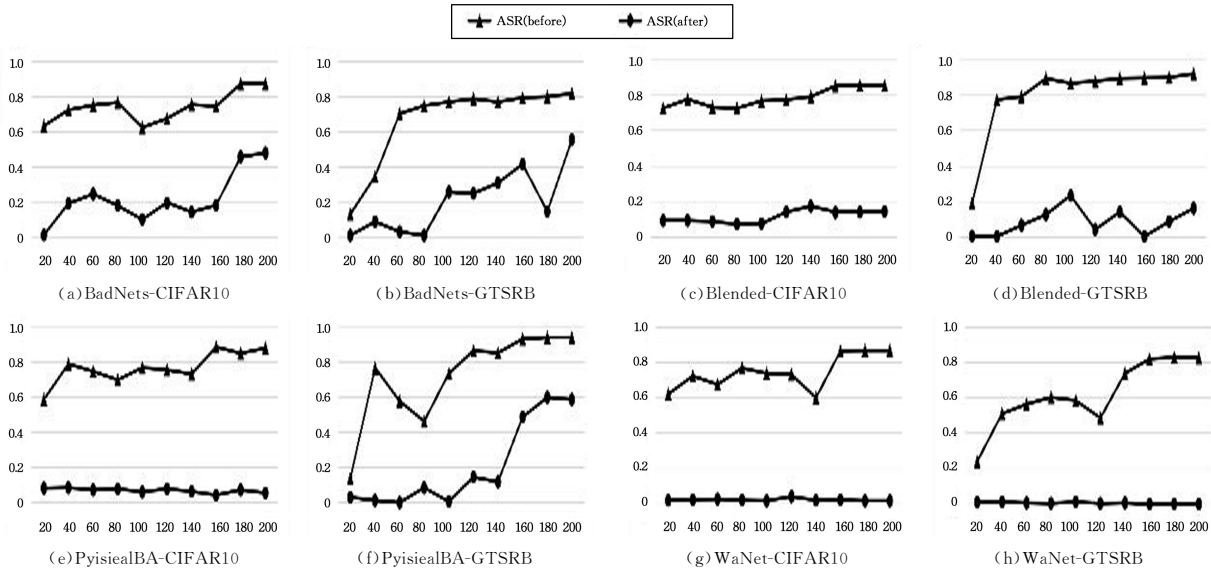

图 9 过滤前后后门植入情况的对比

Fig. 9 Comparison of backdoor implantation before and after filtration

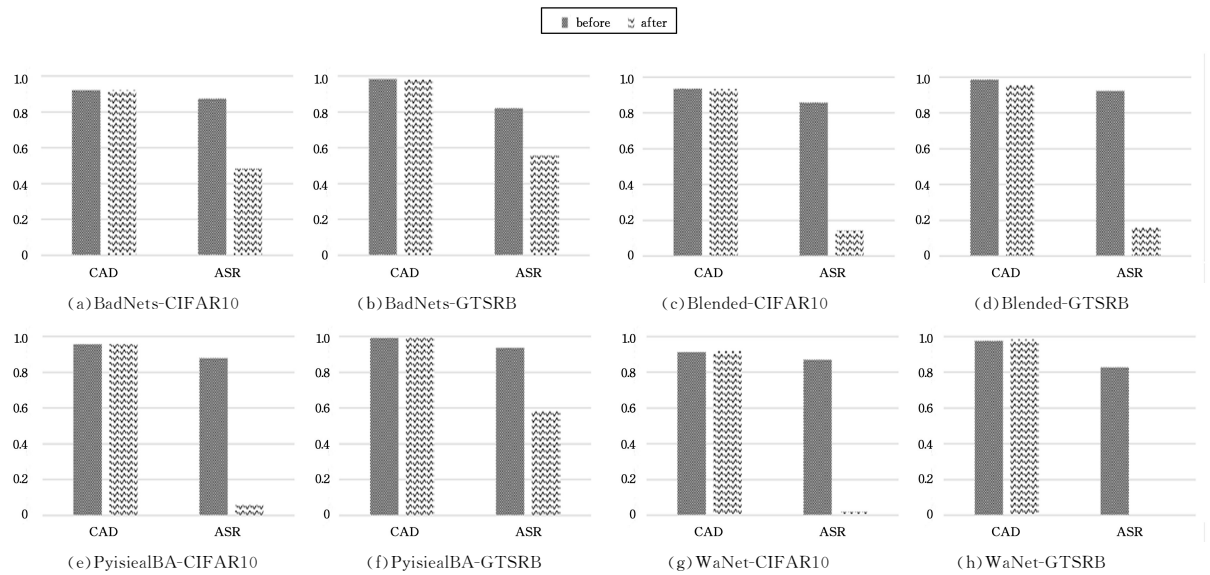

图 10 过滤前后训练结果的对比

Fig. 10 Comparison of training results before and after filtration

4.4 对比实验

本文选取了 SCALE-UP 与 IBD-PSC 两种最新的样本过滤方法进行对比实验，同样在 CIFAR10 与 GTSRB 上针对 Badnets，Blended，PhysicalBA 与 WaNet 进行过滤，并添加了对基于 BadNets 的类不可知攻击的过滤结果作为对照，如表 5、表 6 所列。

表 5 SCALE-UP 和 IBD-PSC 与本文方法在 CIFAR10 上的过滤结果

Table 5 Filtration results of SCALE-UP，IBD-PSC and the proposed method on CIFAR10

(%)

Attacks	*Defenses*	*Recall*	*Precision*	*Accuracy*	*F*1
Badnets (all2one)	IBD-PSC	100	91.48	95.14	95.55
	SCALE-UP	97.72	84.87	89.73	90.84
BadNets	IBD-PSC	17.34	66.47	55.48	27.51
	SCALE-UP	17.34	66.47	55.48	27.51
	ours	97.00	43.93	93.66	59.81
Blended	IBD-PSC	15.19	42.93	48.78	22.44
	SCALE-UP	16.43	49.04	50.75	24.61
	ours	97.25	42.88	93.43	58.78
PhysicalBA	IBD-PSC	10.21	34.77	46.94	15.79
	SCALE-UP	15.19	34.17	44.44	21.03
	ours	97.03	77.79	98.45	86.91
WaNet	IBD-PSC	17.91	31.47	40.93	22.83
	SCALE-UP	11.41	28.76	42.81	16.34
	ours	96.78	52.76	90.88	66.76

表 6 SCALE-UP 和 IBD-PSC 与本文方法在 GTSRB 上的过滤结果

Table 6 Filtration results of SCALE-UP，IBD-PSC and the proposed method on GTSRB

(%)

Attacks	*Defenses*	*Recall*	*Precision*	*Accuracy*	*F*1
Badnets (all2one)	IBD-PSC	1	86.45	92.08	92.73
	SCALE-UP	95.59	80.03	85.72	87.12
BadNets	IBD-PSC	13.81	23.67	37.20	17.44
	SCALE-UP	56.81	46.17	46.87	50.94
	ours	96.34	96.76	99.67	96.55
Blended	IBD-PSC	47.85	75.64	66.97	58.62
	SCALE-UP	35.60	46.79	52.87	40.43
	ours	98.46	58.63	96.45	73.50
PhysicalBA	IBD-PSC	24.03	37.39	44.18	29.25
	SCALE-UP	27.57	37.11	42.15	31.63
	ours	95.52	93.32	99.77	94.41
WaNet	IBD-PSC	25.36	33.60	39.02	28.91
	SCALE-UP	39.49	49.68	54.84	44.00
	ours	97.49	94.54	99.26	96.00

选取 SCALE-UP 与 IBD-PSC 进行对比实验的原因在于其可以体现出类不可知攻击对经验防御方法的冲击：后门逻辑的特殊性使后门样本的特征发生了根本性的变化。这两种防御方法均假设后门样本在参数变换或样本变换中的预测一致性应高于良性样本，但事实上这种差异只稳定地存在于类不可知攻击中。当攻击模式转为类特定攻击时，后门与模型性能强相关的性质对两类样本预测一致性的关系造成了不容忽视的影响，并且不同攻击方法的影响程度各不相同，最终使得针对不同攻击的过滤结果极不稳定，甚至有可能导致错误过滤出更大比例的良性样本。由表 5、表 6 可知，SCALE-UP 与 IBD-PSC 在应对类不可知攻击时有着优秀的表现，甚至能完全过滤后门样本，但在应对类特定攻击时表现糟糕。SCALE-UP 与 IBD-PSC 的部分超参数设置与过滤时良性样本和后门样本的预测一致性均值如表 7、表 8 所列。

表 7 IBD-PSC 与 SCALE-UP 在 CIFAR10 上的预测一致性均值与超参数设置

Table 7 Mean prediction consistency and hyperparameter settings of IBD-PSC and SCALE-UP on CIFAR10

Defenses	Attacks	avg_benign	avg_poison	hyperparam
IBD-PSC	Badnets (all2one)	0.1187	1.0000	$xi=0.4, T=0.9$
	BadNets	0.0887	0.1726	$xi=0.6, T=0.4$
	Blended	0.0887	0.1726	$xi=0.6, T=0.4$
	PhysicalBA	0.1853	0.1430	$xi=0.6, T=0.4$
	WaNet	0.1238	0.1465	$xi=0.6, T=0.3$
SCALE-UP	Badnets (all2one)	0.0394	2.2708	$T=0.5$
	BadNets	0.2572	0.1325	$T=0.2$
	Blended	0.0323	−0.3569	$T=0.2$
	PhysicalBA	0.0217	−0.4745	$T=0.2$
	WaNet	0.0526	−0.4474	$T=0.2$

表 8 IBD-PSC 与 SCALE-UP 在 GTSRB 上的预测一致性均值与超参数设置

Table 8 Mean prediction consistency and hyperparameter settings of IBD-PSC and SCALE-UP on GTSRB

Defenses	Attacks	avg_benign	avg_poison	hyperparam
IBD-PSC	Badnets (all2one)	0.3200	1.0000	$xi=0.4, T=0.9$
	BadNets	0.5551	0.2506	$xi=0.4, T=0.9$
	Blended	0.7687	0.7246	$xi=0.15, T=0.9$
	PhysicalBA	0.3327	0.6099	$xi=0.4, T=0.9$
	WaNet	0.5488	0.5926	$xi=0.4, T=0.9$
SCALE-UP	Badnets (all2one)	−0.0100	1.3300	$T=1$
	BadNets	0.0279	−0.3205	$T=0.2$
	Blended	0.0276	−0.3890	$T=0.2$
	PhysicalBA	0.0167	−0.4909	$T=0.2$
	WaNet	0.0170	0.2030	$T=0.2$

从实验结果中不难看出，相较于类不可知攻击中后门样本与良性样本显著的预测一致性差异，在类特定攻击下，后门样本与良性样本的预测一致性并不具备稳定的大小关系。这使得防御者几乎不可能在黑盒情况下获得稳定且理想的过滤效果，因此才需要基于类特定攻击独有的性质设计针对性的防御方法。

4.5 对类不可知后门样本的过滤结果

本文尝试将所提方法推广至类不可知攻击的场景下，实验设置与结果如表 9 所列。

表 9 针对基于 BadNets 的类不可知攻击的实验设置与过滤结果

Table 9 Experimental setup and filtering results for class-agnostic attacks based on BadNets

	MNIST	CIFAR10
攻击方法	BadNets	BadNets
投毒率	0.05	0.05
迭代轮数	1	4
分区数量	50	7
CDA 均值	0.940254	0.478000
ASR 均值	0.258586	0.903857
Recall	0.981333	0.979200
Accuracy	0.999067	0.785800
Precision	1.000000	0.186785
*F*1-score	0.999533	0.301826

同样以 0.05 的比例进行投毒，在 MNIST 上，即使将训练集拆分为 50 份进行 DPA，结果显示 ASR 依旧稳定低于 CDA，并且 CDA 能保持在 94%左右，因而过滤效果较优，仅一轮迭代就过滤了所有的后门样本。然而，在 CIFAR10 上进行过滤时，由于 ASR 已经稳定超过 CDA，此时去集与留集的作用反转，后门样本集中于 D_{in} 而非 D_{out} 中，最终会在 CDA 下降与干净样本减少的连锁反应下将干净数据“萃取”到 D_{out} 中。但这就并非是 4 次迭代所能完成的工作。

造成这一现象的原因是类不可知攻击的后门逻辑不再依赖于对样本源类的识别。这使得 CDA 与 ASR 不再具备强依赖性，因此性质 1)不再成立；CDA 与 ASR 的鲁棒性也不再有稳定的对比关系，因此性质 3)也不再成立。

结束语 本文方法以类特定攻击下 ASR 对 CDA 的强依赖性为基础，使用 DPA 与投票机制设计了针对类特定攻击的样本过滤方法，在确保稳定过滤后门样本的同时，DPA 也适合进行并行训练，可以有效缩短过滤时间。但针对类特定攻击的防御方法仍亟待探索。

1)本文的防御方法建立在一个标签下的所有干净样本具有相同特征分布的假设下，此时加入投毒样本相当于在目标标签下增加了新的分布，因此，可以通过集成学习的方法分离不同的分布。若一个标签下的干净样本有不同的分布，则在误报率与漏报率之间取得平衡将更加困难。

2)类特定攻击中 ASR 与 CDA 的关系较为稳定，因此可以以相同的模式通过 DPA 与投票法稳定地分离不同分布的样本；类不可知攻击的行为同样会在目标标签下增加新的分布，但 ASR 与 CDA 不再有强关联性。因此，将本文方法应用于防御类不可知攻击时，其模式与效果并不稳定，开销也可能会更高。

3)本文方法基于类特定攻击在训练过程中的固有特征，仅对投毒样本进行过滤，不对触发器做进一步处理。虽然本文方法能稳定降低类特定攻击的威胁性，但在训练 epoch 较高的情况下，漏报的少部分投毒样本仍有可能在模型中植入一个效果较差的后门，第 4 章的部分实验结果证明了这一点，并且在触发器最简单的 BadNets 上表现得最为显著。考虑到样本过滤最终将为防御者提供一个投毒率显著高于原训练集的样本集合，基于过滤结果的触发器检测与消除是一个可行的研究方向。

参 考 文 献

[1] BROWN A, HUH J, CHUNG J S, et al. VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge[J]. arXiv: 2201.04583, 2022.

[2] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: A survey[J]. Science China(Technological Sciences), 2020, 63(10): 1872-1897.

[3] BISONG E. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners[M]. Berkely: Apress, 2019.

[4] YAN B, LAN J, YAN Z. Backdoor attacks against voice recognition systems: A survey[J]. arXiv: 2307.13643, 2023.

[5] LI Y, JIANG Y, LI Z, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5-22.

[6] GAO Y, DOAN B G, ZHANG Z, et al. Backdoor attacks and countermeasures on deep learning: A comprehensive review[J]. arXiv: 2007.10760, 2020.

[7] JAVAHERIPI M, SAMRAGH M, FIELDS G, et al. Cleann: Accelerated trojan shield for embedded neural networks[C]// Proceedings of the 39th International Conference on Computer-Aided Design. 2020: 1-9.

[8] TIAN Z, CUI L, LIANG J, et al. A comprehensive survey on poisoning attacks and countermeasures in machine learning[J]. ACM Computing Surveys, 2022, 55(8): 1-35.

[9] GU T, LIU K, DOLAN-GAVITT B, et al. Evaluating Backdooring Attacks on Deep Neural Networks[J]. IEEE Access, 2019, 7: 47230-47244.

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

[11] DONG Y, YANG X, DENG Z, et al. Black-box detection of backdoor attacks with limited information and data[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16482-16491.

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

[13] GUO J, LI Y, CHEN X, et al. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency[J]. arXiv: 2302.03251, 2023.

[14] HOU L, FENG R, HUA Z, et al. IBD-PSC: Input-level Backdoor Detection via Parameter-oriented Scaling Consistency[J]. arXiv: 2405.09786, 2024.

[15] LEVINE A, FEIZI S. Deep partition aggregation: Provable defense against general poisoning attacks[J]. arXiv: 2006.14768, 2020.

[16] KRIZHEVSKY A. Learning multiple layers of features from tiny images[J/OL]. http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[17] SAADNA Y, BEHLOUL A. An overview of traffic sign detection and classification methods[J]. International Journal of Multimedia Information Retrieval, 2017, 6: 193-210.

[18] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv: 1712.05526, 2017.

[19] LI Y, ZHAI T, JIANG Y, et al. Backdoor attack in the physical world[J]. arXiv: 2104.02361, 2021.

[20] NGUYEN A, TRAN A. Wanet—imperceptible warping-based backdoor attack[J]. arXiv: 2102.10369, 2021.

[21] DOAN K, LAO Y, ZHAO W, et al. Lira: Learnable, imperceptible

ble and robust backdoor attacks[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021:11966-11976.

[22] SOURI H,FOWL L,CHELLAPPA R,et al. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch[J]. Advances in Neural Information Processing Systems,2022,35:19165-19178.

[23] TRAN B,LI J,MADRY A. Spectral Signatures in Backdoor Attacks[J]. arXiv:1811.00636,2018.

[24] HAYASE J,KONG W. SPECTRE:Defending against backdoor attacks using robust covariance estimation[C]// International Conference on Machine Learning. 2021:4129-4139.

[25] ZENG Y,PARK W,MAO Z M,et al. Rethinking the backdoor attacks' triggers:A frequency perspective[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:16473-16481.

[26] HUANG H,MA X,ERFANI S,et al. Distilling cognitive backdoor patterns within an image[J]. arXiv:2301.10908,2023.

[27] AMARNATH C,BALWANI A H,MA K,et al. Tesda:Transform enabled statistical detection of attacks in deep neural networks[J]. arXiv:2110.08447,2021.

[28] CHEN B,CARVALHO W,BARACALDO N,et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv:1811.03728,2018.

[29] LIU G,KHREISHAH A,SHARADGAH F,et al. An adaptive black-box defense against trojan attacks(trojdef)[J]. IEEE Transactions on Neural Networks and Learning Systems,2022,35(4):5367-5381.

[30] CHEN W,WU B,WANG H. Effective backdoor defense by exploiting sensitivity of poisoned samples[J]. Advances in Neural Information Processing Systems,2022,35:9727-9737.

[31] LECUN Y,JACKEL L D,BOTTOU L,et al. Learning algorithms for classification:A comparison on handwritten digit recognition[J]. Neural Networks:the Statistical Mechanics Perspective,1995,261(276):2.

[32] BREIMAN L. Bagging Predictors[J]. Machine Learning,1996,24:123-140.

GUO Jiaming, born in 2000, postgraduate. His main research interests include AI security and offence-defense.

YANG Chao, born in 1982, Ph.D, professor, postgraduate supervisor, is a member of CCF(No. 94791M). His main research interests include information security and computer immunology.

(责任编辑:何杨)