基于自定义后门的触发器样本检测方案

王 尚[1,2], 李 昕[3], 宋永立[3], 苏 铓[1], 付安民[1,2]

[1] 南京理工大学计算机科学与工程学院 南京 中国 210094
[2] 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 中国 100093
[3] 北京计算机技术及应用研究所 北京 中国 100036

摘要 深度学习利用强大的特征表示和学习能力为金融、医疗等多个领域注入新的活力, 但其训练过程存在安全威胁漏洞, 攻击者容易通过操纵训练集或修改模型权重执行主流后门攻击: 数据中毒攻击与模型中毒攻击。两类攻击所产生的后门行为十分隐蔽, 后门模型可以保持干净样本的分类精度, 同时对嵌入攻击者预定义触发器的样本呈现定向误分类。针对干净样本与触发器样本在拟合程度上的区别, 提出一种基于自定义后门行为的触发器样本检测方案 BackDetc, 防御者自定义一种微小触发器并执行数据中毒攻击向模型注入自定义的后门, 接着通过嵌入自定义触发器设计一种输入样本扰动机制, 根据自定义触发器的透明度衡量输入样本的拟合程度, 最终以干净样本的拟合程度为参照设置异常检测的阈值, 进而识别触发器样本, 不仅维持资源受限用户可负担的计算开销, 而且降低了后门防御假设, 能够部署于实际应用中, 成功抵御主流后门攻击以及威胁更大的类可知后门攻击。在 MNIST、CIFAR-10 等分类任务中, BackDetc 对数据中毒攻击与模型中毒攻击的检测成功率均高于目前的触发器样本检测方案, 平均达到 99.8%以上。此外, 论文探究了检测假阳率对检测性能的影响, 并给出了动态调整 BackDetc 检测效果的方法, 能够以 100%的检测成功率抵御所有分类任务中的主流后门攻击。最后, 在 CIFAR-10 任务中实现类可知后门攻击并对比各类触发器样本检测方案, 仅有 BackDetc 成功抵御此类攻击并通过调整假阳率将检测成功率提升至 96.2%。

关键词 深度学习; 后门攻击; 自定义后门; 拟合程度; 触发器样本
中图法分类号 TP391 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.11.03

A Trigger Sample Detection Scheme Based on Custom Backdoor Behaviors

WANG Shang[1,2], LI Xin[3], SONG Yongli[3], SU Mang[1], FU Anmin[1,2]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[3] Beijing Institute of Computer Technology and Application, Beijing 100036, China

Abstract Deep learning leverages powerful feature representation and learning capabilities to breathe new life into various fields such as finance and healthcare, but the training process is vulnerable to security threats, easily introducing mainstream backdoor attacks through manipulating the training data set or modifying model weights, including data poisoning attack and model poison attack. The backdoor implanted by both types of backdoor attacks is great stealthy, the backdoored model can maintain the clean data accuracy, while presenting targeted misclassification for samples embedded with the attacker-specific triggers. This paper proposes a custom backdoor behavior-based trigger samples detection scheme BackDetc, focusing on the essential difference on the fit degree between clean samples and trigger samples. It injects custom backdoors into the model through tiny defender-custom triggers, proposing an input sample perturbation mechanism by embedding these custom triggers. We measure the fit degree of inputs adopting the transparency of the custom trigger, and calculate the threshold of anomaly detection with the fit degree of clean samples as a reference, identifying these samples with attacker-specific triggers. In this way, BackDetc not only holds the affordable overhead for resource limited users, but reduces the strength of backdoor defense assumption, being deployed in various real-world applications and being effective for mainstream backdoor attacks as well as more threatening source-specific backdoor attacks. In experiments, the BackDetc is deployed on MNIST, CIFAR-10 classification tasks, outperforming other existing trigger samples detection schemes on detection success rate when facing data poisoning attack and model poison attack, with an average of over 99.8%. Then, the influence of the detection false positive rate is explored on the detection performance, giving the capability of dynamically adjusting the detection effect of BackDetc, displaying 100% detection success rate on all tasks when encountering two mainstream backdoor attacks. Meanwhile, in the CIFAR-10 task, a source-specific backdoor attack is

通讯作者: 付安民, 博士, 教授, Email: fuam@njust.edu.cn。
本课题得到国家自然科学基金(No. 62072239), 江苏省自然科学基金(No. BK20211192), 信息安全国家重点实验室开放基金(No. 2021-MS-07)资助。
收稿日期: 2022-06-20; 修改日期: 2022-08-04; 定稿日期: 2022-09-07

implemented to evaluate various trigger samples detection schemes, only BackDetc successfully resists such the attack and increases the detection success rate to 96.2% by adjusting the false positive rate.

Key words deep learning; backdoor attack; customize backdoor; fit degree; trigger samples

1 引言

海量的数据与丰富的应用场景促进了深度学习技术的蓬勃发展, 赋能众多应用领域, 如智慧医疗、智慧金融和智慧交通等[1-3]。但深度神经网络训练过程具有两大弱点: 数据依赖性与模型不可解释性, 导致人工智能系统存在若干安全威胁[4], 如投毒攻击、对抗样本攻击与后门攻击。投毒攻击[5]通过修改、删除或者注入精心设计的数据来破坏模型, 最终阻止训练过程收敛或降低模型精度。对抗样本攻击[6]作用于模型预测阶段, 为数据精心设计肉眼不可见的扰动, 从而误导模型的预测效果。而后门攻击[7-8]预先定义触发器并将其嵌入部分训练数据中, 通过训练使模型对不含触发器的干净样本执行正常预测行为。而携带触发器的样本将激活后门行为, 即定向分类为攻击者预先选定的标签。

后门攻击过程中, 攻击者首先确定目标标签与触发器, 构造少量携带触发器的数据, 并将其标签修改为目标类别, 通过多轮训练来学习触发器与目标标签之间的强连接, 同时学习干净样本与真实标签之间的特征映射, 不仅维持干净样本的分类精度, 而且使得触发器样本具有极高的攻击成功率, 说明后门攻击相较于其他安全攻击更加隐蔽且危险。而且, 攻击者通常采用肉眼不可见的触发器向模型植入后门, 强化其威胁能力[9]。现实世界中, 后门攻击已经威胁到公共安全[10-11], 例如, 嵌入后门的交通标志识别系统可能导致自动驾驶的汽车错误识别携带触发器图案的交通标识。

根据攻击者构造触发器的思路可将后门攻击分为数据中毒攻击[12]与模型中毒攻击[7], 由此可以延伸出大量后门攻击方案, 因此本文重点检测此两类主流后门攻击下的触发器样本。此外, 目前出现一种后门变体攻击, 对深度学习模型产生更大的威胁, 即类可知后门攻击[13], 嵌入此变体后门的模型仅对攻击者指定类别的触发器样本产生定向误分类, 对于干净样本与非指定类别的触发器样本仍保持良性的预测行为。现有的后门防御方案认为触发器仅包含非鲁棒的过拟合特征, 而类可知后门攻击将触发器特征与攻击者指定类别的特征相结合而产生后门行为[14], 成功绕过此防御假设以及大多数防御方案。

为应对后门攻击威胁, 研究人员针对后门攻击的三要素: 触发器、受损神经元以及两者间的连接, 提出多种后门防御手段。针对触发器特性, 部分研究致力于探索触发器样本与干净样本之间的差异[15-16], 认为触发器存在过拟合的非鲁棒特征, 进而从数据集或输入样本中过滤触发器样本。而一些研究者根据受损模型中的后门特性检测或消除受损神经元, 包括模型重建[17]、模型诊断[18]等。针对触发器与受损神经元间的连接, 部分研究者通过输入样本预处理[19]或模型剪枝微调[20]切断此连接, 使得触发器无法激活后门。但是后门攻击存在若干切入点, 包括毒化数据集、外包模型训练任务与迁移学习等。上述方案由于防御假设受限, 仅能够防御特定场景下的主流后门攻击, 对其他场景鲁棒性较差, 而且无法抵御威胁性更强的类可知后门攻击。同时, 这些方案操作复杂, 资源开销不稳定, 无法有效部署于资源受限的用户端[21]。因此, 目前亟需一种轻量级、易部署且能够抵御多种后门攻击的防御方案。

我们关注后门模型对干净样本与触发器样本的预测行为, 从特征维度分析两者差别。为提高触发器样本的攻击成功率, 攻击者利用中毒数据集进行训练, 迫使模型将触发器含有的特征与目标标签建立强映射关系, 即触发器特征存在过拟合特性[22]。而对于干净样本, 模型根据图像内容的关键特征分析其隶属各类别的概率, 且用户为提高模型泛化能力, 有意控制干净样本的拟合程度, 均导致其中的良性特征拟合程度偏弱。因此, 当样本携带触发器时, 其中的良性特征与触发器特征同时输入至后门模型, 后者会优先激活模型的过拟合行为从而分类为目标标签。因此, 触发器样本与干净样本在特征的拟合程度上存在明显区别, 前者的抗干扰能力远高于后者, 因此可以通过噪声干扰输入样本预测结果的难易程度设计触发器样本检测机制[16]。

因此, 本文从扰动方式入手, 设计有效的扰动来衡量输入样本的抗干扰能力, 即拟合程度, 参照干净数据的拟合程度识别触发器样本。常规扰动机制容易影响攻击者指定触发器的效果, 即无法准确衡量触发器样本的拟合程度, 此现象在类可知后门攻击[14]中尤为明显。为提高后门防御方案的鲁棒性, 提出一种基于自定义后门行为的触发器样本检测方案 BackDetc, 防御者自定义一个微小的触发器, 向模型注入自定义后门, 此时能够利用自定义触发器

执行干扰过程, 根据触发器的透明度衡量输入样本的拟合程度, 最终统计干净数据的拟合程度集合以确定检测阈值, 进而识别携带原始触发器的样本。需注意, 自定义触发器由防御者设计, 而原始触发器由攻击者构造, 为避免混淆, 下文中触发器样本仅表示携带原始触发器的样本。

该方案关注触发器样本与干净样本的本质区别, 利用自定义后门行为执行干扰机制, 不仅对原始触发器性质鲁棒, 而且保持有限的资源开销, 特别适合于资源有限且安全需求较高的应用场景, 如物联网设备。本文的主要贡献具体如下:

1) 设计了一种基于自定义后门行为的输入样本干扰机制, 采用微小且不影响视觉的自定义触发器向模型注入自定义后门, 以自定义触发器的透明度计算输入样本的抗干扰能力, 即拟合程度, 并探究干净样本与触发器样本在预测过程的本质区别。

2) 基于输入样本干扰机制, 提出一种基于自定义后门行为的触发器样本检测方案 BackDetc, 收集干净数据的拟合程度来确定检测阈值, 进而识别触发器样本。相较于当前的触发器样本检测方案, BackDetc 不仅提升了触发器样本的检测成功率, 而且保持资源有限用户可负担的计算开销。同时, 其扰动机制仅依赖自定义后门行为, 成功抵御对其他检测方案有效的类可知后门攻击。

3) 分别在 MNIST 与 CIFAR-10 数据集上执行各类后门攻击, 并部署 BackDetc 检测各后门攻击中的触发器样本, 其中主流后门攻击下的检测成功率平均达到 99.8%以上。同时利用检测假阳率进行消融实验, 动态调整 BackDetc 的检测性能, 将类可知后门攻击下的检测成功率提升至 96.2%。本方案操作简洁, 在线阶段仅需少量预测步骤即可检测任意输入样本, MNIST 任务中仅需 3.9ms 即可完成一次检测。

2 背景知识

2.1 后门攻击

深度学习训练过程可分为若干阶段, 出于效率与模型精度考虑, 用户倾向于将部分阶段外包给算力强大、数据量丰富的第三方, 如采用第三方收集的数据、外包训练任务给第三方以及微调第三方发布的开源模型[22]。而恶意第三方为破坏模型完整性, 作为攻击者从其中一点切入来注入后门。本方案对 3 种攻击场景均有效, 因此我们选取攻击能力最大的情况, 即用户将训练数据与训练任务交付至第 3 方[23], 注入后门并将模型返回给用户, 如图 1 所示。训练过程中, 攻击者拥有用户的全部知识, 包括训练数据集、训练算法、深度学习模型的结构与内部参数。因此, 攻击者可以修改训练数据及相应的标签, 控制训练过程, 甚至直接修改模型的内部参数。此假设最大化攻击能力, 在白盒情况下部署后门攻击, 具有出色的攻击成功率, 也使防御更具挑战性。下文以上述攻击假设为基础考虑目前主流的后门攻击, 通过建立非鲁棒特征与受损神经元间的强映射来实现后门行为, 主要包括数据中毒攻击与模型中毒攻击。

数据中毒攻击[12]。攻击者随机选择任意图案作为触发器, 将其嵌入部分训练数据中并修改它们的标签, 以此构造中毒数据集, 接着通过训练将触发器特征与目标标签建立强连接。其攻击过程如图 1(a)所示, 用户将训练数据与训练任务交由第三方以期获得高精度的深度学习模型。恶意第三方作为攻击者首先确定触发器图案与目标类别, 随机抽取部分训练数据, 于固定位置嵌入触发图案并将这些数据的标签修改为目标类别, 即获得中毒数据, 与干净数据混合后得到中毒数据集。后续通过多轮训练成功引入后门, 并将受损模型返回至用户。由于受损模型在干净样本上保持较高的精度, 用户验证模型精度后直接部署后门模型, 严重威胁深度学习应用的安全性。具体地, 当干净样本携带触发器时, 后门模型根据过拟合的触发器特征将其分类为目标类别, 而忽略样本中存在的良性特征。

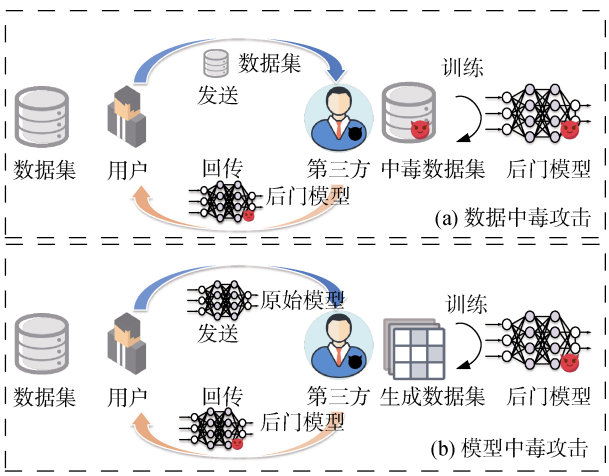


图 1 主流后门攻击框架

Figure 1 The architecture of mainstream backdoor attacks

模型中毒攻击[7]。与前者思路不同, 攻击者无需操纵训练数据集, 选定特定神经元集合后利用逆向工程生成触发器与部分数据, 按照数据中毒攻击构造中毒数据集并执行模型训练过程, 以此将后门嵌入模型并返回给用户。其攻击过程如图 1(b)所示, 攻

击者选择一组神经元与触发器初始形状, 以扩大十倍选定神经元集合的激活值为目标更新触发区域, 迭代生成触发器并建立异常连接, 即木马触发器。然后, 利用逆向工程构造各类数据, 按照数据中毒攻击类似的方法将木马触发器嵌入部分生成数据中, 对应数据的标签修改为目标类别。最终, 以上述的异常强连接为桥梁将触发器与目标标签建立强映射关系。因此, 木马触发器可视为更强的过拟合特征, 推理阶段中携带木马触发器的干净样本直接激活目标神经元, 进而映射为目标类别。

2.2 类可知后门攻击

目前主流的后门攻击均属于源不可知后门攻击, 对应的受感染模型对任意嵌入触发器的样本均呈现后门行为, 即产生定向误分类, 其后门效果与输入样本的内容无关。尽管主流后门攻击能够实现接近100%的攻击成功率, 但是多数后门防御方案基于源不可知特性成功抵御此类后门攻击[24]。同时在实际应用中, 主流后门攻击存在误报的缺陷。例如, 攻击者以特制的眼镜为触发器将后门注入人脸识别系统, 期望佩戴特制眼镜的攻击者激活后门。而其余良性用户佩戴类似的眼镜将极大概率产生相同的后门效果, 这将引起相关部门的重视。

类可知后门攻击成功解决主流后门攻击存在的限制。一方面, 嵌入类可知后门的模型仅对攻击者指定类别的触发器样本产生定向误分类, 对干净样本与非指定类别的触发器样本保持良性的预测行为, 其后门效果与输入样本的内容相关[25]。因此, 类可知后门攻击将触发器特征与指定类别的特征相结合, 成功绕过基于源不可知特性的后门防御方案[26]。另一方面, 实际应用中, 非指定类别的样本携带与触发器相似的配件难以产生定向误分类, 保证了后门效果的隐蔽性。为构造类可知的后门效果, 与主流后门攻击中的训练数据集不同, 此变体攻击选择指定类别, 向其中部分数据中添加触发器并修改标签, 以生成中毒数据。对于非指定类别, 向其中部分数据中添加触发器并维持真实标签, 得到恢复数据。将中毒数据、恢复数据与干净数据混合后执行训练过程, 此模型即嵌入类可知后门。

综上所述, 类可知后门攻击给深度学习模型的安全性造成了更加严重的威胁, 而且目前缺少有效的防御手段。

2.3 后门检测方案

尽管目前已存在大量后门防御方案, 但它们鲁棒性普遍较差, 仅对一部分后门攻击有效, 甚至存在无法忽略的局限性。Liu 等人[20]忽略模型是否存在后门, 采用干净数据集对神经网络进行剪枝微调操作, 虽抑制后门效果却降低了模型精度。而文献[27]采用人工脑刺激技术搜索对特定类别表现异常激活值的神经元, 对触发器性质敏感且计算开销极大, 难以落地于实际应用场景。因此, 本文主要关注鲁棒性较强的触发器样本检测方案, 并选取其中效果显著的 3 种方案作为 BackDetc 的参照, 以下展开详细描述。

Neural Cleanse[18]包含一种在线输入样本检测方案。防御者将模型的后门视为迁移至目标类别的捷径, 通过逆向工程生成所有类别的捷径, 即候选触发器。然后, 采用异常检测技术分析所有候选触发器的 L 范数, 依据真实触发器微小原则判断模型是否存在后门。最后, 通过逆向触发器激活的神经元检测输入样本是否携带触发器。此方案需要构造所有类别的捷径, 消耗大量计算资源且对触发器性质敏感, 如位置、尺寸与透明度等。

SentiNet[28]是一种输入样本检测方案。防御者认为触发器对分类结果具有重要的影响, 因此通过目标检测技术确定输入图像中影响分类的重要区域, 并判断其中是否包含触发器。然后, 将此区域嵌入干净样本中得到第一类检测数据, 同时消除干净样本中此区域所在的位置得到第二类检测数据, 当前者误分类率与后者分类精度均高时, 可判定所选区域存在触发器, 即可识别触发器样本。此方案需要利用干净数据集构造元分类器, 操作复杂且计算开销极大, 极易受触发器性质影响。

STRIP[16]是一种在线触发器样本检测方案, 通过触发器样本难以扰动的特点完成检测。防御者生成若干输入样本的副本, 将等量随机的干净样本与之混合, 若预测结果较分散, 说明输入样本不携带触发器。反之, 说明输入样本存在难以扰动的特征, 即判定为触发器样本。此方案虽与 BackDetc 类似, 但随机样本扰动机制可控性差且极大概率影响原始触发器的效果, 无法防御类可知后门攻击与输入可知后门攻击[29]。

3 种防御方案均存在一些弱点, 目前亟需一种轻量级、操作简洁且能够抵御主流后门攻击与类可知后门攻击的触发器样本检测方案。因此, 本文提出一种基于自定义后门行为的触发器样本检测方案 BackDetc, 能够满足上述所有特性。

3 基于自定义后门行为的触发器样本检测方案

3.1 防御假设

防御能力。用户作为防御者, 尽管将训练任务外

包至第三方, 依旧可以访问干净数据集与模型内部参数。一般情况下, 用户具有微调模型的计算资源, 倾向于利用本地数据集微调外包模型。理论上, 用户具备插入自定义后门的能力与资源。同时, 本方案仅调整模型部署时的预测行为, 并未修改外包模型的内部参数。

防御目标。防御者具有以下 3 个目标:

1) 对于任意模型, 维持干净样本的分类精度;

2) 对于数据中毒攻击、模型中毒攻击与类可知后门攻击, 保持极高的检测成功率;

3) 操作简洁且维持资源受限用户可承担的计算开销。

3.2 触发器样本检测方案概述

用户获得第三方训练完成的外包模型后, 依次从离线与在线两个阶段完成触发器样本检测方案, 如图 2 所示。离线阶段包括两个步骤: 将自定义的后门注入外包模型副本, 收集干净样本的拟合程度集合。其目的是确定干净样本拟合程度的上界, 以此获得触发器样本在拟合程度上的检测阈值。在线阶段, 用户仅需要计算输入样本的拟合程度, 将其与检测阈值比较, 即可判断此样本是否携带触发器。

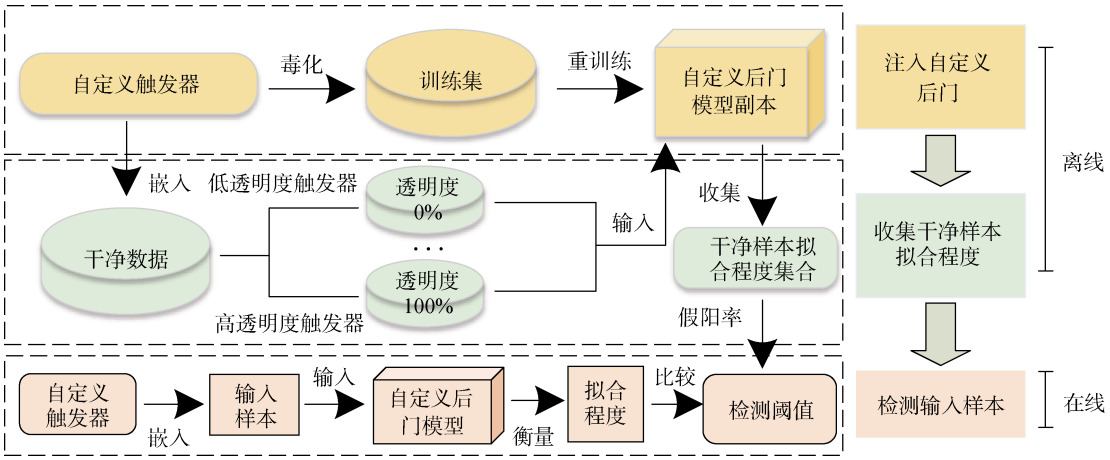


图 2 触发器样本检测方案架构图

Figure 2 The architecture of trigger sample detection scheme

离线阶段, 用户首先构造携带自定义后门的外包模型副本。具体地, 随机选择微小且不影响视觉效果的像素块作为自定义的触发器, 同时随机选择训练集中的部分数据, 将用户选择的触发器嵌入这些数据的固定位置, 修改相应数据的标签为用户预选的类别。通过在模型副本上执行少量轮次的训练, 成功将自定义后门注入外包模型副本。其次, 利用自定义触发器易于激活自定义后门的特点, 设计一种对原始触发器特性影响小的输入样本干扰机制, 以此计算任意输入样本的抗干扰程度, 即拟合程度。具体地, 针对某输入样本, 为其生成若干副本, 依次将低透明度到高透明度的自定义触发器嵌入这些副本并采用嵌入自定义后门的模型预测其结果。当标签由用户预选的类别改变为其它类别时, 说明此时输入样本的特征强度高于对应透明度的自定义触发器, 即为输入样本的抗干扰能力或拟合程度。用户随机选择部分干净数据并统计相应的拟合程度集合, 最终计算触发器样本的检测阈值。

在线阶段, 用户根据安全需求选择合适的检测假阳率, 排除部分拟合程度较高的干净样本, 选择其余拟合程度集合中的最大值作为检测阈值。此时, 用户仅需要基于自定义触发器的输入样本干扰机制计算当前样本的拟合程度, 将其与检测阈值比较, 即可判断此样本是否携带触发器。

本方案从干净样本与触发器样本在预测阶段的本质区别入手, 降低防御假设的强度, 通过检测过拟合的特征识别触发器样本, 对多种后门攻击具有鲁棒性。同时, 此方案资源开销集中于离线阶段, 而且自定义后门注入以及干净样本拟合程度收集所需的计算开销在用户可负担范围内, 适用于实际的防御场景。

4 方案细节

针对基于自定义后门行为的触发器样本检测方案 BackDetc, 本节对其中关键的技术以及流程进行详细描述, 包括自定义后门注入策略, 输入样本拟合程度测量算法以及触发器样本检测流程。

4.1 自定义后门注入策略

图 3 显示了自定义后门注入外包模型的过程。用户首先为外包模型生成一个副本, 模拟攻击者执行数据中毒攻击, 通过预选的简易触发器构造中毒数据集, 并以此重新训练模型副本, 即得到嵌入自定义后门的模型副本, 同时不影响外包模型本身。具体来说, 用户首先确定微小且不影响图像视觉的触发图案 m, 以及目标类别 t。然后, 从非目标类中随机采样 5%的数据并于角落位置添加自定义的触发器 m, 同时将其标签更新为目标类别 t, 混入训练数据即得到中毒数据集 D_p。最后, 用户生成外包模型的副本 M, 利用交叉熵损失函数 L 在中毒数据集上训练训练少量轮次, 其目标函数如下:

$$\arg\min_{\theta} \sum_{(x_i,y_i)\in D_p} L(M(x_i;\theta),y_i)$$

因为数据中毒攻击并不影响干净样本的预测精度, 携带自定义后门的模型副本在主分类任务上具有较高的准确率, 而当预测嵌入自定义触发器的输入样本时, 自定义后门被激活, 即预测结果均为目标类别 t。同时本文证明, 外包模型若嵌入后门, 自定义后门的注入行为并不影响原始的后门效果, 即触发器样本的拟合程度仍高于干净样本, 保证了 BackDetc 检测方案的有效性。

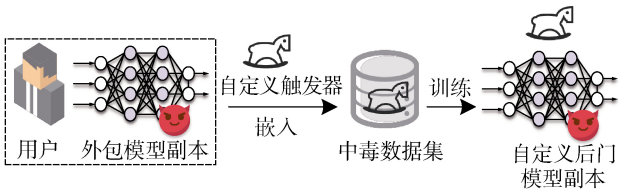


图 3 自定义后门插入步骤

Figure 3 The procedure of customized backdoor insertion

4.2 输入样本拟合程度测量算法

后门攻击过程中, 攻击者利用数据中毒攻击或模型中毒攻击毒化训练数据集或操控模型参数, 迫使模型将触发器特征与目标标签 a 建立强映射关系, 即触发器特征存在过拟合特性。触发器样本由于非良性特征拟合程度过高, 导致其以极高的置信度分类为目标标签 a。而干净样本依据其中关键的良性特征完成预测, 如模型将含有车轮、车灯与方向盘等元素的图像分类为汽车。为保证模型的泛化能力, 良性特征的拟合程度通常不高。因此, 本文引入自定义后门行为实现干扰机制, 以输入样本的抗干扰能力作为其拟合程度, 进而区分干净样本与触发器样本。

我们提出一种输入样本拟合程度测量算法, 利用自定义触发器的透明度表示输入样本的拟合程度, 如图 4 所示。针对一个输入样本, 首先获得外包模型预测标签 t_0, 从其余类别中随机选取一种类别 t_1 作为自定义后门的目标标签, 利用后门注入策略构建一个嵌入自定义后门的模型副本 M_1, 其对应的自定义触发器为 m_1。测算过程中, 用户生成 n 份输入样本的副本并分别添加触发器 m_{11}, m_{12},…, m_{1n}, 其中 m_1 的透明度依次递增, 从不透明的触发器 m_{11} 到全透明的触发器 m_{1n}, 相应的自定义后门效果逐渐降低, 同样自定义触发器的干扰能力随之下降。最终, 将所有样本依次输入模型副本 M_1 获得相应的预测标签。其中预测类别由 t_1 转变为 t_0 时, 说明输入样本中的特征强度高于此时的自定义触发器, 即将 m_1 对应的透明度定义为输入样本的拟合程度。

在携带自定义后门的模型副本 M_1 中, 自定义触发器 m_1 极易激活后门行为, 将携带 m_1 的输入样本定向误分类为目标标签 t_1。而依次增加 m_1 的透明度将导致自定义后门的效果逐渐降低, 即表现出原本的预测标签 t_0。若输入一个攻击者设计的触发器样本, 由于触发器特征拟合程度过高, m_1 不透明时即输出 t_0 为预测标签。而输入一个干净样本, m_1 透明度较高时才输出 t_0 为预测标签。因此, 本文依据输入样本对自定义触发器的抵抗能力作为其特征的拟合程度, 从而区分干净样本与触发器样本。下文详细描述输入样本拟合程度测量算法。

算法 1.输入样本拟合程度测量算法.

输入: 外包模型 M、输入样本 x;

输出: 输入样本拟合程度 q;

①生成外包模型副本 M_1, 并确定 x 的预测标签 t_0;

②确定自定义的触发器 m_1, 目标标签 t_1;

③构造中毒数据集并训练模型副本 M_1;

④生成 n 个输入样本副本 $x_1, x_2,\ldots, x_n$;

⑤按透明度递增序列依次添加 n 个触发器 m_{11}、$m_{12},\ldots, m_{1n}$;

REPEAT

⑥利用 M_1 预测 $x_i+ m_{1i}$, 其标签为 $label_i$;

UNTIL $label_i= t_1$ 且 $label_{i+1}=t_0$;

RETURN m_{1i} 的透明度 q.

4.3 在线触发器样本检测方案 BackDetc 构建

基于输入样本拟合程度测量算法, 我们设计在线触发器样本检测方案的具体步骤。用户首先通过微小且不影响视觉效果的触发器构造携带自定义后门的模型副本, 然后随机选择部分干净数据, 利用输入样本拟合程度测量算法收集良性拟合程度的集

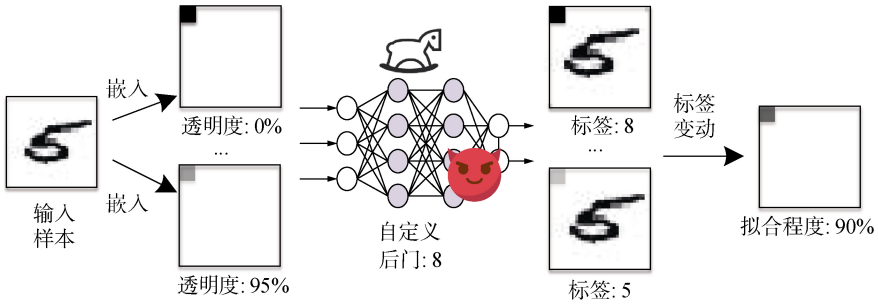


图 4　输入样本拟合程度测量步骤

Figure 4　The measure steps of input fit degree

合。为保证检测方案不影响干净样本的分类精度, 理想情况下, 拟合程度集合中选择最大值作为检测阈值, 针对任意输入样本, 计算其拟合程度并与阈值比较。若高于阈值, 则将其视为触发器样本, 且外包模型必然存在后门。

考虑到输入样本的原始类别与自定义类别偶然重合的可能, 本方案设置两个具有不同目标类别的自定义后门, 其中必然存在与原始类别不相同的自定义后门, 因此输入样本需要分别在两个不同的自定义后门下进行检测, 存在超出阈值的情况即可判定为触发器样本。为提供灵活的检测能力, 用户通过修改检测假阳率来动态更新检测阈值, 即初始认定少量干净样本的拟合程度存在异常, 排除它们后选择剩余拟合程度集合中的最大值作为检测阈值, 以模型精度在假阳率范围内降低为代价提高触发器样本的检测成功率。用户具体按照以下步骤执行输入样本检测过程:

1) 用户获得外包训练模型 M 并拷贝其参数获得两个模型副本, 随机选择两个类别作为两个自定义后门的目标类别, 同时生成两种微小且不影响视觉的自定义触发器, 利用自定义后门注入策略构造两个中毒数据集, 以此训练出两个携带不同自定义后门的模型副本;

2) 用户从测试集中选择部分干净数据, 按照输入样本拟合程度测量算法分别在两个模型副本中获得所有干净样本的拟合程度集合;

3) 用户按照防御需求设置合理的假阳率, 并按照递减序列确定对应的拟合程度, 进而作为触发器样本的检测阈值;

4) 针对任意输入样本, 用户在两个模型副本中分别计算其拟合程度。若其中任意拟合程度高于对应的阈值, 即可判定此输入样本含有触发器, 模型 M 存在第三方注入的后门, 其余情况则判定此输入样本为干净样本。

BackDetc 的核心是触发器特征存在过拟合特性, 引入自定义后门计算输入样本的拟合程度, 以干净数据的拟合程度为参考筛除触发器样本, 即设置合理的拟合程度作为检测阈值, 具体过程为算法 2。此过程对应的计算开销与模型微调相近, 资源受限用户在本地即可实现离线阶段的部署。后续, 在线阶段仅需少量预测行为即可检测某输入样本是否嵌入触发器, 操作简洁且易于部署, 如算法 3 所示。

此外, 在 BackDetc 构建过程中, 用户需要于图像中隐蔽位置嵌入微小的自定义触发器, 如角落。因此自定义触发器以极低概率影响输入样本的关键内容或攻击者嵌入的触发区域, 导致自定义后门具有较强的可控性, 以此削弱干扰机制对原始触发器的影响, 从而保证 BackDetc 对原始触发器性质的鲁棒性以及对多数后门攻击的有效性。理论上, 存在过拟合行为的后门攻击均无法逃避此检测。

算法 2. 触发器样本检测阈值构造算法.

输入: 外包模型 M、假阳率 f、干净数据集 D_c;

输出: 模型副本 M_1 与 M_2、检测阈值 T_1 与 T_2.

①随机初始化目标类别 t_1 与 t_2, 微小触发器 m_1 与 m_2;

②利用 m_1 与 t_1 执行自定义后门注入策略, 获得模型副本 M_1;

③利用 m_2 与 t_2 执行自定义后门注入策略, 获得模型副本 M_2;

④利用 M_1、m_1 与 t_1 计算 D_c 所有样本的拟合程度, 其集合为 Q_1;

⑤利用 M_2、m_2 与 t_2 计算 D_c 所有样本的拟合程度, 其集合为 Q_2;

⑥选取 Q_1 与 Q_2 排序第 f 的拟合程度作为检测阈值 T_1 与 T_2;

RETURN 模型副本 M_1 与 M_2、检测阈值 T_1 与 T_2.

算法 3. 在线触发器样本检测算法.

输入: 模型副本 M_1 与 M_2、自定义触发器 m_1 与 m_2、目标类别 t_1 与 t_2、检测阈值 T_1 与 T_2、任意输入样本 x;

输出: 触发器性质 *flag*.

①利用 M_1、m_1 与 t_1 计算 x 的拟合程度 q_1;

②利用 M_2、m_2 与 t_2 计算 x 的拟合程度 q_2;

③若 $q_1 < T_1$ 且 $q_2 < T_2$, 则 *flag=clean*;

④否则 *flag=trigger*;

RETURN 触发器性质 *flag*.

5 实验分析

本节按照攻击假设分别在 MNIST 与 CIFAR-10 数据集中实现数据中毒攻击与模型中毒攻击, 然后对四种受感染的外包模型执行基于自定义后门行为的触发器样本检测方案 BackDetc, 选择部分干净样本与触发器样本进行检测并记录防御效果, 探讨检测假阳率对本方案中触发器样本检测成功率的影响。后续引入三种效果显著的触发器样本检测方案[16,18,28], 从防御效果、资源开销方面与 BackDetc 展开对比, 并比较各方案对类可知后门攻击的鲁棒性。最后, 通过上述结果分析本方案的特性与适用场景。

5.1 实验设置

我们采用图 5(a—c)所示触发器完成后门攻击操作, 均以类别 0 为攻击目标, 随机修改 5%训练数据或生成木马触发器, 通过数据中毒攻击与模型中毒攻击分别将后门注入 MNIST 与 CIFAR-10 分类模型[30]。与文献[18]中攻击设置相同, MNIST 任务采用图 5(a)(b)作为触发器, 分别实现数据中毒攻击与模型中毒攻击。而 CIFAR-10 任务选用图 5(a)(c)作为触发器实现两类主流的后门攻击。以 MNIST 任务为例, 两种后门攻击产生的恶意行为如图 6(a)(b)所示, 任意一种后门攻击中, 嵌入触发器的样本均分类为攻击者预先定义的类别 0。

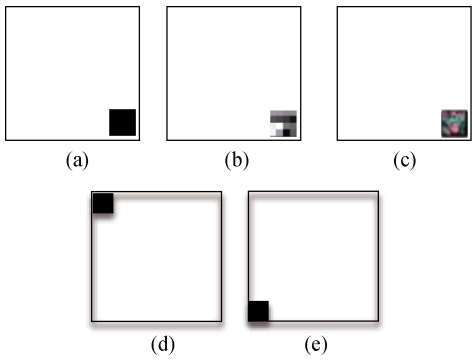


图 5 后门攻击的触发图案

Figure 5 The trigger patterns of backdoor attacks

实验中, 攻击者硬件环境为: Intel i5-9300H (2.40GHz)CPU 和 64GB 内存; 用户硬件环境为: Intel i5-6500CPU 和 16GB 内存。两者的软件执行环境均为: Windows10 操作系统、Python 3.6.2 和 PyTorch 1.14.0。

表 1 中描述了 MNIST 与 CIFAR-10 数据集的信息摘要。其中 MNIST 具有 10 个类别, 包含 70000 张灰度手写数字图像, 其中 60000 张属于训练集, 10000 张属于测试集, 采用 2 层卷积 2 层全连接的神经网络[31]实现分类。而 CIFAR-10 数据集中具有 10 种通用物体, 包含 60000 张彩色图像, 其中 50000 张属于训练集, 10000 张属于测试集。考虑到任务的复杂性, CIFAR-10 采用 Rsnet18 模型进行实现[32]。

表 1 数据集摘要

Table 1 Dataset Summary

数据集	类别	模型	训练/测试
MNIST	10	2Conv+2Dense	60000/10000
CIFAR-10	10	Resnet18	50000/10000

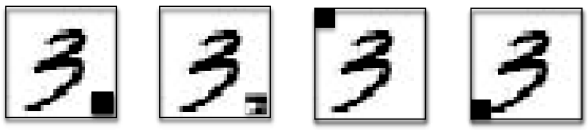


图 6 触发器样本的预测结果

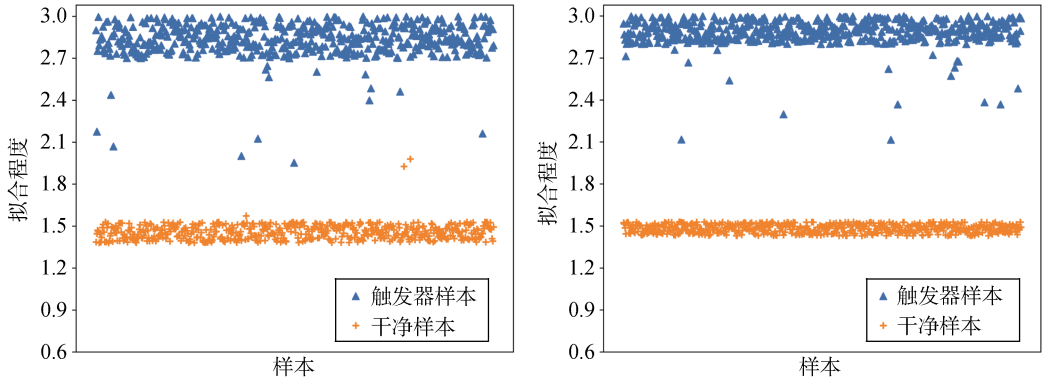
Figure 6 The prediction results of trigger samples

5.2 自定义后门注入策略

针对某个受感染的外包模型, 我们分别采用图 5(d)(e)中自定义的微小触发图案执行自定义后门注入策略, 其中图 5(d)与类别 8 建立强映射, 而图 5(e)与类别 9 建立强映射。针对某个模型副本, 首先随机选择 5%训练数据添加自定义触发器, 并将其标签修改为相应的目标类别, 以此构造中毒数据集, 然后通过少量轮次训练即可将自定义后门注入模型副本。为保证 BackDetc 方案对任意样本均有效, 需要消除输入样本的原始类别与自定义后门目标类别重合的偶然性, 因此生成两个携带不同自定义后门的模型副本。后续, 输入样本在任意自定义后门检测机制下表现异常, 即可判为触发器样本。

实验证明, 经过两至三轮训练, 模型副本即可成功嵌入自定义后门, 其资源开销与模型微调接近, 用户在本地即可完成此过程。我们采用 1000 份干净样本与 1000 份嵌入自定义触发器的样本计算此模型副本的分类精度以及自定义后门的攻击成功率。所有任务中任意攻击下的模型副本, 它们不仅维持干净样本的分类精度, 而且平均实现 99%以上的自定义后门攻击成功率。以 MNIST 任务为例, 如图 6(c)(d)所示, 嵌

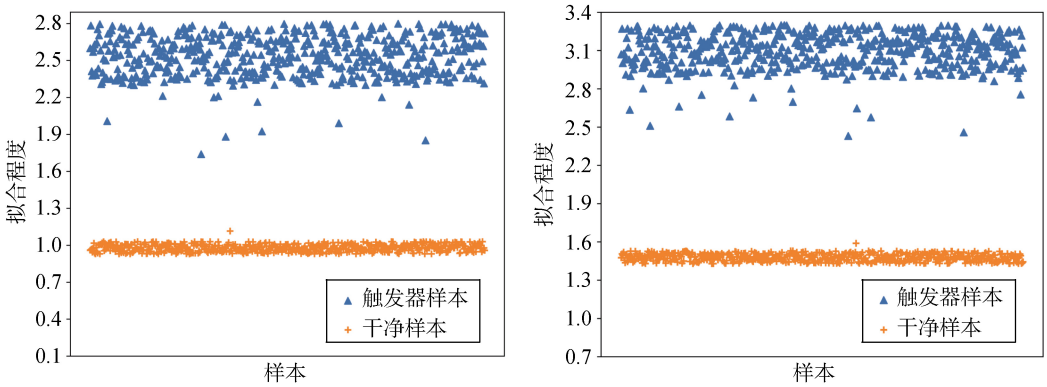
入自定义触发器的样本极易激活相应的自定义后门, 产生定向误分类。此外, 我们计算攻击者引入的后门所造成的攻击成功率, 携带原始触发器的样本仍维持100%的攻击成功率, 证明自定义后门注入过程并不影响已存在的后门行为, 即携带原始触发器的样本仍存在过拟合行为, 即可被 BackDetc 方案检测。



(a) 数据中毒攻击　　(b) 模型中毒攻击

图 7　MNIST 后门攻击下样本的拟合程度

Figure 7　The fit degree of samples on backdoor MNIST



(a) 数据中毒攻击　　(b) 模型中毒攻击

图 8　CIFAR-10 后门攻击下样本的拟合程度

Figure 8　The fit degree of samples on backdoor CIFAR-10

5.3　触发器样本检测方案 BackDetc

本节主要验证 BackDetc 对两类主流后门攻击的防御效果。针对某个嵌入后门的分类模型, 我们均匀选择 1000 份干净样本, 并制作 1000 份触发器样本。首先, 从测试集中随机选取 20%的干净数据, 分别输入两个携带不同自定义后门的模型副本中, 利用触发器样本检测阈值构造算法获得两个检测阈值, 接着通过在线触发器样本检测方案识别携带原始触发器的样本。

为方便理解, 下文以映射为类别 8 的后门模型副本为例绘制 1000 份干净样本与 1000 份触发器样本的拟合程度散点情况。在 MNIST 分类任务中, 数据中毒攻击检测效果如图 7(a)所示, 其中触发器样本的拟合程度显著高于干净样本。若以干净样本中最高拟合程度为检测阈值, 即检测假阳率设置为 0%, 则触发器样本的检测成功率可达 99.8%。而模型中毒攻击的检测效果如图 7(b)所示, 其中触发器样本与干净样本在拟合程度上存在分界, 在检测假阳率设置为 0%的条件下可实现 100%的检测成功率。模型中毒攻击下生成的触发器直接影响受损神经元, 相较于随机产生的触发器具有更高的拟合程度, 导致图 7(b)中 2 种样本拟合程度的分隔增大。对于干净样本, 仍输入至外包模型获得预测结果, 因此本方案并不影响分类精度, 保持在 98%以上。

同样, 我们针对 CIFAR-10 下的两类后门攻击部署 BackDetc 方案, 以映射为类别 9 的后门模型副本为例绘制 1000 份干净样本与 1000 份触发器样本的拟合程度分布情况。图 8(a)(b)分别显示了 BackDetc 对数据中毒攻击与模型中毒攻击的检测效果, 2 种攻击下触发器样本与干净样本在拟合程度上均存在分界, 因此选择干净样本中最高拟合程度为检测阈值即可实现 100%检测成功率。与 MNIST 任务相比, 此

分类任务更加复杂，干净样本中的良性特征拟合程度偏低，而触发器特征仍然保持过拟合特性，因此干净样本与触发器样本在拟合程度上相差更大，容易实现 100%的检测效果。对于干净样本，将其输入至外包模型，仍保持与干净模型相近的分类精度，维持在 88%以上。

此外，检测假阳率是影响防御效果的关键因素，提高假阳率会降低检测阈值，能够提升检测效果，而过高的假阳率直接影响分类精度，因此后续将讨论不同假阳率对分类精度以及 BackDetc 检测效果的影响。针对 4 种攻击情况，分别在检测假阳率为 0%、0.5%、1%与 2%的情况下部署本方案，结果如图 9 所示。以 MNIST 任务中数据中毒攻击为例，触发器样本检测成功率呈现增长趋势。当假阳率高于 0.5%时，其检测成功率均达到 100%。对于其他 3 种攻击情况，同样在假阳率高于 0.5%时实现 100%的检测成功率。两种分类任务中，当假阳率不超过 0.5%时，外包模型的分类精度平均变化为 0.2%，可忽略不计。同时，当假阳率为 0%时，平均检测成功率可达 99.9%。因此，后续实验中假阳率可设置为 0%或 0.5%，以此权衡检测效果与分类精度。

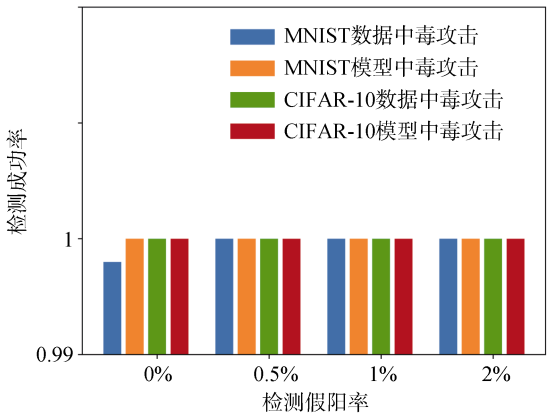


图 9　检测假阳率差异实验

Figure 9　The difference of detection false positive rate

5.4　触发器样本检测方案对比分析

为进一步评估BackDetc方案的性能与资源开销，我们与当前效果显著的 3 种触发器样本检测方案进行实验对比，其中包括 STRIP[16]、Neural Cleanse[18]与 SentiNet[28]，最大化 3 种方案的防御能力，利用充足的计算能力与丰富的数据资源部署各方案。本实验中，Neural Cleanse 在测试集上利用逆向工程为每个类生成一个候选触发器，通过异常检测方案识别真实触发器，后续以此为检测依据判断输入样本是否携带触发器。而 SentiNet 通过模型解释与目标检测技术捕获输入样本的关键区域，统计测试集中干净样本对应的关键区域特性并训练元分类器，从而识别触发器样本。STRIP 通过随机叠加 100 个干净样本实现干扰机制，利用预测结果的熵值作为输入样本的拟合程度，并设置合理的假阳率以确定检测阈值。与本方案相似，STRIP 设置为 0%的假阳率。针对本文四种攻击场景，我们随机生成 2000 份触发器样本，部署以上四种输入样本检测方案来记录相应的检测成功率，其结果如表 3 所示。

由此表可以看出，BackDetc 在主流后门攻击情况下均具有绝对的优势，其检测成功率高于另外三种方案。Neural Cleanse 对模型中毒攻击的防御效果较差，其中逆向工程倾向于搜索微小且规整的触发器，无法完全还原木马触发器，导致识别触发器样本时出现偏差。SentiNet 需要统计大量干净样本关键区域的特性，提高了元分类器的鲁棒性，对两类后门攻击均可防御，但仍存在一些触发器样本位于分类边界，造成偏低的检测成功率。STRIP 表现整体超出另外两种方案，但叠加干净样本的扰动机制会降低原始触发器的效果，导致某些触发器样本的熵值低于阈值，即出现少量逃逸检测的情况。而自定义触发器足够小，并不影响输入样本中原始触发器的效果，因此 BackDetc 表现出更强的检测效果。

表 2　触发器样本检测方案资源开销

Table 2　The overhead of trigger sample detection schemes

防御方案	资源开销(ms)	
	离线	在线
STRIP[16]	1070	10.4
Neural Cleanse[18]	169190	0.4
SentiNet[28]	65270	17.6
BackDetc	4590	3.9

为进一步评估各方案的资源开销，我们以 MNIST 下数据中毒攻击为例记录四种方案的资源开销。各方案包括离线部署与在线检测两个阶段，其结果如表 2 所示。Neural Cleanse 为所有类别构造候选触发器，离线阶段计算资源消耗巨大；SentiNet 的计算开销包括目标检测、模型解释、大量干净样本关键区域特性分析以及元分类器的训练，在离线阶段同样消耗大量计算资源。而 STRIP 仅需要一些模型预测操作即可完成离线部署与在线检测操作。本方案离线阶段包括自定义后门注入以及干净样本拟合程度收集，其开销虽略高于 STRIP，但远低于其余两

种方案, 容易部署于资源受限的用户端。此外, BackDetc 在线阶段所需的模型预测操作低于 STRIP, 将单次输入样本检测时间降低至 3.9ms。综上所述, 本方案检测效果显著高于当前的方案, 且两阶段计算开销满足资源受限用户的需求, 可落地于真实的应用场景。

表 3 触发器样本检测方案对比

Table 3 The comparison of trigger sample detection schemes (%)

防御方案	MNIST 检测成功率		CIFAR-10 检测成功率	
	数据中毒攻击	模型中毒攻击	数据中毒攻击	模型中毒攻击
STRIP[16]	98.1	98.7	99.1	99.5
Neural Cleanse[18]	96.4	92.8	95.7	91.3
SentiNet[28]	94.3	95.1	94.3	94.8
BackDetc	99.8	100	100	100

5.5 类可知后门攻击检测效果

后门变体攻击进一步威胁深度学习安全, 如类可知后门攻击, 将攻击目标锁定为攻击者指定的类别, 即源类。因此, 仅源类样本才可以激活后门行为, 而模型对携带触发器的非源类样本保持良性预测行为。Tang 等人[14]详细分析了类可知后门攻击的威胁能力, 可成功绕过 Neural Cleanse 与 STRIP 方案。后者虽与本文思路相似, 但是采用叠加随机样本的方式构造扰动无法准确衡量输入样本的拟合程度。因为类可知后门攻击下原始触发器仅对特定类别起作用, 叠加随机样本极大可能削弱原始触发器的后门效果, 导致某些触发器样本的熵值接近干净样本。而 SentiNet 基于类不可知特性实现检测机制, 仅针主流后门攻击有效, 同样无法抵御类可知后门攻击。而本方案单独考虑输入样本的拟合程度, 且扰动机制不影响触发器效果。即使针对特定类别的触发器, 嵌入源类样本后, 激活类可知后门, 此恶意行为的拟合程度极大概率高于干净样本的预测行为。下文采用 CIFAR-10 实现类可知后门攻击, 并部署 BackDetc 检测携带触发器的源类样本。

首先, 以图 4(a)为触发器, 类别 5(狗)为源类且类别 0(飞机)为目标类, 根据文献[14]中步骤生成中毒数据与恢复数据, 实现类可知后门攻击。由图 10 可知, 类可知后门将嵌入触发器的源类样本分类为目标类别 0, 而此效应对其他类别无效。经验证, 该后门模型在干净样本上的分类精度为 87.53%, 对于嵌入触发器的源类样本的攻击成功率高达 97.8%, 说明此后门变体攻击兼顾隐蔽性与破坏性。检测过程中, 构造携带自定义后门的模型副本, 随机选择 1000 份干净样本与 1000 份源类中的触发器样本并绘制拟合程度分布情况。

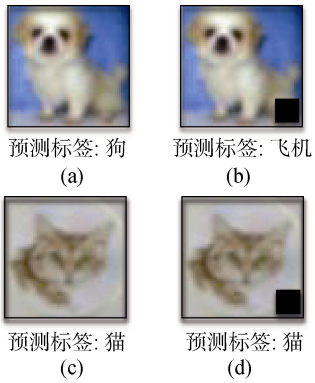


图 10 类可知后门攻击样例

Figure 10 The examples of source-specific backdoor attack

图 11 表示干净样本与携带触发器的源类样本在拟合程度上存在明显差异, 即源类样本中触发器仍存在过拟合特性, 因此 BackDetc 可以抵御类可知后门攻击。当检测假阳率为 0%时, 本方案的检测成功率达到 88.3%。尽管一些触发器样本可以逃逸检测, 本方案仍可通过调整假阳率来降低检测阈值, 即获得更高的检测成功率。我们绘制出 0%、0.5%、1%与 2%假阳率对应的检测效果, 如图 12 所示, 牺牲 2%模型精度即可将检测成功率提升至 96.2%, 说明 BackDetc 的鲁棒性远高于其他 3 种检测方案, 可有效抵御类可知后门攻击。

6 讨论

多数后门防御方案容易受触发器性质、模型复杂度的影响[33-34], 而本方案采用微小且不影响视觉的自定义触发器注入自定义后门, 保证干扰机制的独立性, 只依据原始触发器存在的过拟合特性完成检测, 理论上对模型结构与触发器性质均不敏感。

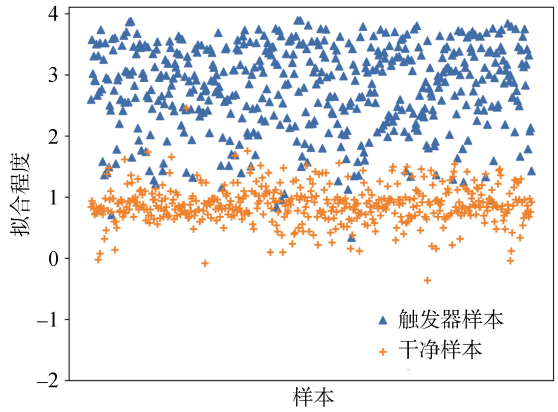


图 11 类可知后门下样本的拟合程度

Figure 11 The fit degree of samples on source-specific backdoor

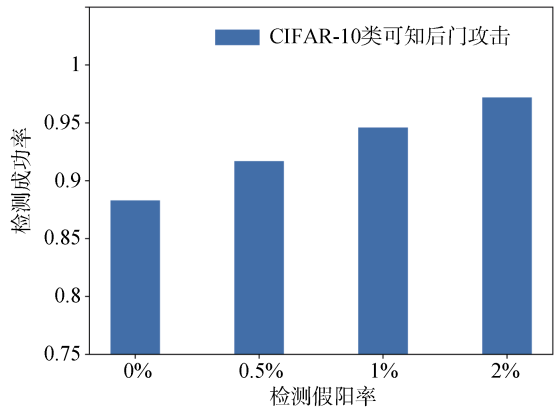


图 12 不同检测假阳率下类可知后门的检测效果

Figure 12 The detection effect of the source-specific backdoor under different detection false positive rates

本节以图像右下角的像素块作为原始触发器，在 CIFAR-10 任务中实现五种数据中毒攻击，探究 BackDetc 在不同触发器尺寸下的防御效果，原始触发器占输入图像的比例包括 1%、5%、10%、15%与 20%，如图 13 所示。考虑到 STRIP 方案与本方案的相似性，我们同时记录两者对触发器尺寸的鲁棒性，其假阳率均设置为 0%，相应的检测效果如图 14 所示。随着触发器尺寸增加，STRIP 方案对触发器样本的检测成功率略微下降，最低仍高于 96%，而本方案的检测成功率维持在 99%以上。由此说明两种方案对触发器尺寸均不敏感，其中 BackDetc 的鲁棒性高于 STRIP。我们认为，通过随机叠加样本的扰动方式更容易遮挡大尺寸触发器，影响攻击者注入的后门效果，即扰动本身削弱了原始触发器的抗干扰能力，造成部分触发器样本逃逸 STRIP 检测的情况。而本方案的扰动方式仅在图像边缘或角落添加微小的自定义触发器，难以遮挡原始触发器，更精确地计算输入样本的拟合程度，因此对原始触发器尺寸表现出更强的鲁棒性。

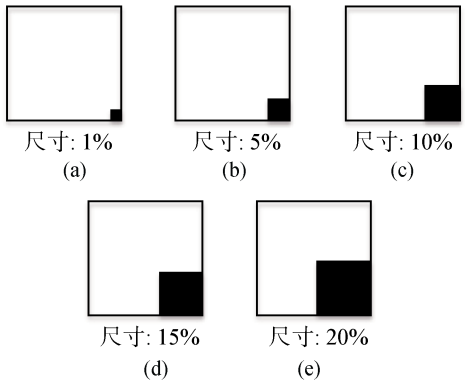


图 13 不同尺寸的原始触发器

Figure 13 The different attacker-specific triggers with different sizes

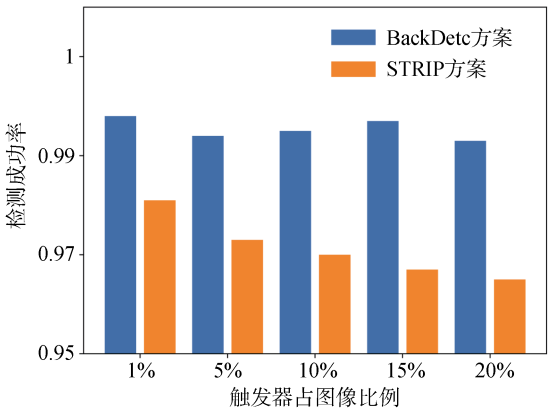


图 14 不同触发器尺寸下 BackDetc 与 STRIP 对比

Figure 14 The comparison of BackDetc and STRIP under different trigger sizes

此外，通过上述实验我们发现后门攻击的效果越强，相应触发器的特征拟合程度越高，导致干净样本与触发器样本在拟合程度上呈现更加显著的差别[35]。由此可提出一种绕过本方案的新型后门攻击，迫使触发器特征与良性特征拟合程度接近。显然，这使得触发器样本无法以极高成功率完成定向误分类，增强了自适应后门攻击的难度与实用性。因此，攻击者期望绕过本方案，必须承受后门攻击成功率下降的代价，以此保证了防御的鲁棒性。

7 结论

本文提出一种基于自定义后门行为的触发器样本检测方案 BackDetc，使用自定义触发器的透明度衡量输入样本的拟合程度，通过触发器特征存在过拟合的特点区分触发器样本与干净样本。本方案引入独立的自定义后门，保证干扰机制不影响原始触发器的效果，以此提高检测的鲁棒性。与目前效果显

著的触发器样本检测方案相比, BackDetc 不仅对主流后门攻击具有99.8%以上的检测成功率, 而且解决了类可知后门攻击的威胁, 将检测成功率提升至96.2%。同时, 本方案在离线与在线阶段均保持较低的计算开销, 可部署于资源受限的用户端, 且整体步骤操作简洁, 适用于真实应用场景。在未来工作中, 我们将进一步降低自定义后门对已存在后门的影响, 以便设计更加鲁棒的触发器样本检测方案。

参考文献

[1] Chen J Y, Zou J F, Su M M, et al. Poisoning Attack and Defense on Deep Learning Model: A Survey[J]. *Journal of Cyber Security*, 2020, 5(4): 14-29.
(陈晋音, 邹健飞, 苏蒙蒙, 等. 深度学习模型的中毒攻击与防御综述[J]. *信息安全学报*, 2020, 5(4): 14-29.)

[2] Fu A M, Zhang X L, Xiong N X, et al. VFL: A Verifiable Federated Learning with Privacy-Preserving for Big Data in Industrial IoT[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(5): 3316-3326.

[3] Tan Z W, Zhang L F. Survey on Privacy Preserving Techniques for Machine Learning[J]. *Journal of Software*, 2020, 31(7): 2127-2156.
(谭作文, 张连福. 机器学习隐私保护研究综述[J]. *软件学报*, 2020, 31(7): 2127-2156.)

[4] Zhou C Y, Chen D W, Wang S, et al. Research and Challenge of Distributed Deep Learning Privacy and Security Attack[J]. *Journal of Computer Research and Development*, 2021, 58(5): 927-943.
(周纯毅, 陈大卫, 王尚, 等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. *计算机研究与发展*, 2021, 58(5): 927-943.)

[5] Xiao H, Xiao H, Eckert C. Adversarial Label Flips Attack on Support Vector Machines[J]. *Frontiers in Artificial Intelligence and Applications*, 2012, 242: 870-875.

[6] 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.

[7] Liu Y Q, Ma S Q, Aafer Y, et al. Trojaning attack on neural networks[C/OL]. *Network and Distributed System Security Symp*, https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf. 2018.

[8] Chen D W, Fu A M, Zhou C Y, et al. Federated Learning Backdoor Attack Scheme Based on Generative Adversarial Network[J]. *Journal of Computer Research and Development*, 2021, 58(11): 2364-2373.
(陈大卫, 付安民, 周纯毅, 等. 基于生成式对抗网络的联邦学习后门攻击方案[J]. *计算机研究与发展*, 2021, 58(11): 2364-2373.)

[9] Doan K, Lao Y J, Li P. Backdoor Attack with Imperceptible Input and Latent Modification[C]. *The Annual Conf on Neural Information Processing Systems*, 2021: 18944-18957.

[10] Wenger E, Passananti J, Bhagoji A N, et al. Backdoor attacks against deep learning systems in the physical world[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6202-6211.

[11] Feng Y, Ma B T, Zhang J, et al. FIBA: Frequency-Injection Based Backdoor Attack in Medical Image Analysis[EB/OL]. 2021: arXiv: 2112.01148. https://arxiv.org/abs/2112.01148

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

[13] Chen Z Z, Wang S, Fu A M, et al. LinkBreaker: Breaking the Backdoor-Trigger Link in DNNS via Neurons Consistency Check[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2000-2014.

[14] Tang D, Wang X F, Tang H X, et al. Demon in the Variant: Statistical Analysis of DNNS for Robust Backdoor Contamination Detection[EB/OL]. 2019: arXiv: 1908.00686. https://arxiv.org/abs/1908.00686

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

[16] 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.

[17] Zhao P, Chen P Y, Das P, et al.Bridging mode connectivity in Loss Landscapes and Adversarial Robustness[C/OL]. *The 8th International Conference on Learning Representations*, https://openreview.net/group?id=ICLR.cc/2020/Conference. 2020.

[18] 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.

[19] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems[C]. *ACSAC '20: Annual Computer Security Applications Conference*, 2020: 897-912.

[20] Liu K, Dolan-Gavitt B, Garg S. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks[M]. Research in Attacks, Intrusions, and Defenses. Cham: Springer International Publishing, 2018: 273-294.

[21] 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.)

[22] Li Y M, Jiang Y, Li Z F, et al. Backdoor Learning: A Survey[EB/OL]. 2020: arXiv: 2007.08745. https://arxiv.org/abs/2007.08745

[23] Tao G H, Liu Y Q, Shen G Y, et al. Model orthogonalization: class distance hardening in neural networks for better security[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 1372-1389.

[24] Wang S, Gao Y S, Fu A M, et al. CASSOCK: Viable Backdoor Attacks Against DNN in the Wall of Source-Specific Backdoor Defences[EB/OL]. 2022: arXiv: 2206.00145. https://arxiv.org/abs/2206.00145

[25] Gao Y S, Doan B G, Zhang Z, et al. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review[EB/OL]. 2020: arXiv: 2007.10760. https://arxiv.org/abs/2007.10760

[26] Tan T J L, Shokri R, Communication N A B T, et al. Bypassing backdoor detection algorithms in deep learning[C]. *2020 IEEE*

European Symposium on Security and Privacy, 2020: 175-183.

[27] 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.

[28] Chou E, Tramèr F, Pellegrino G, et al. SentiNet: detecting localized universal attacks against deep learning systems[C]. *2020 IEEE Security and Privacy Workshops*, 2020: 48-54.

[29] Salem A, Wen R, Backes M, et al. Dynamic Backdoor Attacks Against Machine Learning Models[EB/OL]. 2020: arXiv: 2003.03675. https://arxiv.org/abs/2003.03675

[30] Ning R, Li J, Xin C S, 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.

[31] Hou B Y, Gao J Q, Guo X J, et al. Mitigating the Backdoor Attack by Federated Filters for Industrial IoT Applications[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(5): 3562-3571.

[32] 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.

[33] Nguyen A, Tran A. Input-Aware Dynamic Backdoor Attack[EB/OL]. 2020: arXiv: 2010.08138. https://arxiv.org/abs/2010.08138

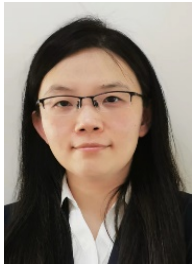
[34] Li Y Z, Li Y M, Wu B Y, et al. Invisible backdoor attack with sample-specific triggers[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2022: 16443-16452.

[35] Lin J Y, Xu L, Liu Y Q, et al. Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 113-131.

王尚 于 2020 年在南京师范大学计算机科学与技术(师范)专业获得学士学位。现在南京理工大学学校网络空间安全专业攻读硕士学位。研究领域为人工智能安全、联邦学习。研究兴趣包括: 隐私保护、后门攻击。Email: shihewang1998@163.com

李昕 于 2009 年在中国航天工业总公司第二研究院计算机应用技术专业获得硕士学位。现任北京计算机技术及应用研究所高级工程师, 主要研究领域为工业控制系统、工业信息安全。研究兴趣包括: 工业控制系统、工业信息安全。Email: lx83@live.cn

宋永立 于 2013 年在北京航空航天大学机械工程专业获得硕士学位。现任北京计算机技术及应用研究所高级工程师, 主要研究领域为工业控制系统、工业信息安全。研究兴趣包括: 工业控制系统、工业信息安全。Email: songyongli2006@126.com



苏铓 于2014年在西安电子科技大学密码学专业获得博士学位。现任南京理工大学副教授。研究领域为网络空间安全。研究兴趣: 云计算安全、隐私保护。Email: sumang@njust.edu.cn

付安民 于 2011 年在西安电子科技大学信息安全专业获得博士学位。现任南京理工大学教授、博士生导师。研究领域为网络空间安全。研究兴趣包括: 网络与系统安全、数据安全与隐私保护。Email: fuam@njust.edu.cn