

# 预训练模型辅助的后门样本自过滤防御方法

刘琦, 张天行, 陆小锋, 吴汉舟, 毛建华, 孙广玲

(上海大学通信与信息工程学院, 上海 200444)

**摘要:** 深度神经网络由于其出色的性能, 被广泛地部署在各种环境下执行不同的任务, 与此同时它的安全性变得越来越重要。近年来, 后门攻击作为一种新型的攻击方式, 对用户构成严重威胁。在训练阶段, 攻击者对少量样本添加特定后门模式并标记为目标类以学习后门模型。后门模型可以以很高的概率将加入后门模式的测试样本识别为目标类, 同时不影响正常样本的识别。用户通常无法掌握后门的先验信息, 因此很难察觉后门攻击的存在。该文提出一种预训练模型辅助的后门样本自过滤方法, 以防御后门攻击, 包括目标类检测与后门样本自过滤两个部分。在第一部分, 利用预训练模型提取样本特征, 采用k近邻算法进行目标类检测; 在第二部分, 从非目标类样本中学习部分分类模型, 之后多次执行“后门样本过滤”与“模型学习”的交替计算, 在较好过滤后门样本的同时, 也得到了完整的良性模型。

**关键词:** 深度神经网络; 后门攻击; 预训练模型; k近邻; 自过滤

中图分类号: TP309.2

文献标识码: A

文章编号: 1673-629X(2023)01-0121-09

doi:10.3969/j.issn.1673-629X.2023.01.019

## Self-filtering of Backdoor Samples by Aid of Pre-trained Model

LIU Qi, ZHANG Tian-xing, LU Xiao-feng, WU Han-zhou, MAO Jian-hua, SUN Guang-ling

(School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China)

**Abstract:** While deep neural networks (DNNs) have been widely deployed in various environments due to their excellent performances, serious security threats emerge accordingly. As a new type of attack in recent years, the backdoor attack composes one of the most serious threats which users are suffered from. The backdoor attack occurs when the attacker changes pixels in a minor amount of training images locally or globally using specific backdoor pattern called ‘trigger’, and also specifies the target label. Tested sample injected the same trigger will be classified into the target label with a high probability regardless of its ground truth, and the benign sample classification performance will not be impacted. Users usually have no prior knowledge about the backdoor attack, thereby the backdoor attack is not easy to be exposed. We propose a backdoor sample self-filtering by the aid of pre-trained model to defend against backdoor attack which contains two components: target class detection and backdoor samples’ self-filtering. At the first component, by using certain pre-trained model, feature representation is extracted for each sample, and then the k-nearest neighbor algorithm (kNN) is used to detect the target class. At the second component, a partial model is learned from the non-target class samples first, and then an iterative and alternative procedure of backdoor sample filtering and benign sample learning is conducted. Finally, not only backdoor samples are filtered out but a complete benign model is obtained as well.

**Key words:** deep neural networks; backdoor attack; pre-trained model; kNN; self-filtering

## 0 引言

随着人工智能的不断进步, 神经网络也实现了飞速发展。神经网络模型已经在计算机视觉<sup>[1]</sup>、语音识别<sup>[2]</sup>、自动驾驶<sup>[3]</sup>等多个领域都表现出优越的性能。神经网络的大规模应用为用户带来便利的同时, 也衍生出一些安全风险<sup>[4]</sup>。神经网络模型缺乏可解释性和透明性是其得到更广泛推广的重要阻碍, 尽管许多学者通过设计特定的结构<sup>[5]</sup>来探究深度

模型的本质或是通过可视化手段<sup>[6]</sup>来强化模型的可解释性, 但目前仍无法对深度模型的行为给出合理的解释。攻击者可以通过这一缺陷对模型造成危害, 对抗攻击首先证明了模型的脆弱性, 在测试阶段, 只要将精心设计的对抗扰动附加在测试数据上, 就会使网络产生错误分类<sup>[7]</sup>。

另一方面, 深度学习模型的优越性很大程度依赖于大量的训练数据和计算资源。受到实际条件的制

收稿日期: 2022-02-24

修回日期: 2022-06-27

基金项目: 上海市科委科技创新行动计划项目(21511102605); 国家自然科学基金项目(61902235)

作者简介: 刘琦(1996-), 男, 硕士研究生, 研究方向为后门攻击与防御; 通讯作者: 孙广玲, 博士, 副教授, 研究方向为深度模型攻击与防御、可解释性。

约,个体用户通常没有用于训练大型复杂模型的计算资源<sup>[8]</sup>,也无法获取高质量的训练数据集<sup>[9]</sup>,因此会将训练任务外包至第三方计算平台以降低计算成本。这意味着用户将在一定程度上失去对数据集和训练过程的控制权。后门攻击就是一种主要的存在于训练阶段的安全风险<sup>[10]</sup>,Gu等人<sup>[11]</sup>通过毒害深度学习模型的训练数据集,首次将后门触发植入到深度模型中。具体而言,攻击者将训练集中的部分样本进行中毒(向样本中添加触发器并修改其标签为目标类别),使模型中的某些神经元对特定的后门模式产生强烈响应,以达到篡改模型的目的。添加的后门触发器可以小到只有一个像素点<sup>[12]</sup>,或是物理世界中的墨镜<sup>[13]</sup>。在测试阶段,当输入带有特定触发模式的后门样本时,这些后门样本会激活模型中的后门,使模型产生定向的错误分类,分类到攻击者指定的目标类别,而模型在干净样本的分类工作中仍保持良好的性能,这种特性使得用户很难检测模型中后门的存在。以交通标志识别任务为例,攻击者在一些“停车”标识图像上添加后门触发器,并将这些样本的标签更改为“限速”标识,这会导致用户在使用通过该数据集训练得到的模型完成自动驾驶任务时产生错误判断,造成难以挽回的损失。

如上文所述,后门攻击是一种存在于训练阶段的更加复杂、破坏力更大的潜在性威胁。攻击者可以通过以下三种场景植入后门:(1)攻击者仅提供后门数据集,模型训练过程完全由用户掌握;(2)攻击者不仅提供后门数据集,还掌握训练过程;(3)攻击者向网络开源数据库中提供的公共模型中嵌入后门,通过网络将后门模型分发给用户。这些丰富的应用场景及其特有的隐蔽性为后门防御任务带来了严峻的挑战。近期的许多研究试图在不同场景下减轻后门攻击带来的威胁,包括用户在无法控制训练过程的前提下,诊断一个模型是否携带后门触发<sup>[14]</sup>并利用剪枝等操作移除对后门模式有特殊响应的异常神经元<sup>[15]</sup>。然而,这些方法都不能彻底地清除后门模型中的后门触发,且会对模型在正常样本分类任务的性能造成影响。当用户使用攻击者提供的后门样本自行训练模型时,通过基于样本过滤的防御方法<sup>[16]</sup>过滤后门样本,但在现实场景中该方法也存在一定的局限性。

该文提出的防御方法适用于以下场景:攻击者是后门数据集的提供方,拥有访问和操纵数据集的权限,攻击者向训练数据集中的少量样本注入特定的后门触发模式并将标签修改为指定标签后交付用户,训练任务由用户完成,攻击者无法控制训练过程。首先,借助预训练模型和k近邻算法判断给定数据集是否注入了后门样本,并确定目标类别;然后,从非目标类别样本

中学习一个部分分类模型,后续基于迭代机制学习完整分类模型,而完整分类模型的学习与更新需要通过过滤掉后门样本的数据集实现,因此采用“后门样本过滤”与“模型学习”交替计算的框架。无论是部分分类模型还是完整分类模型,都依赖于模型自身的分类能力实现后门样本的自过滤。同时,在模型学习的迭代机制中,对正常样本的分类性能和对后门样本的过滤能力都获得提高。

该文的主要贡献如下:

(1)借助于预训练模型,采用k近邻算法检测目标类,并依赖于模型自身的分类能力,实现后门样本的自过滤。

(2)在模型迭代学习与更新的机制中,逐步增强对正常样本的分类性能,也提高对后门样本的过滤能力。

(3)面向多个图像数据集的分类任务和多个攻击模式,以端到端和微调两种方式展开实验验证,并与其他方法进行比较,表明了所提方法的可行性和具备的优势。

## 1 相关工作

### 1.1 后门攻击方法

Gu等人<sup>[11]</sup>率先发现并定义了后门攻击,为这一方向的研究提供了重要参考。此后基于后门触发的后门攻击从触发器是否可见的角度可以分为两类:可见后门攻击和不可见后门攻击。以BadNets为代表的可见后门攻击通过在部分良性图像 $x$ 上覆盖一个局部的像素块(后门触发器)得到中毒图像 $x'$ ,并与目标标签 $y_t$ 相关联得到后门样本 $(x', y_t)$ ,用中毒后的数据集训练网络模型,得到的模型将带有后门能力。虽然可见后门攻击通常能达到较强的攻击效果,但触发模式的隐蔽性较差。Chen等人<sup>[12]</sup>首先讨论了后门模式不可见性的要求,在这之后,一系列致力于探究更隐蔽的后门触发器和更先进的后门添加方式的攻击方法被提出。Zhu等人<sup>[17]</sup>提出了一种图像隐写技术,可以将信息通过隐写的方式嵌入图像,将此技术应用于后门攻击中,通过向图像中写入特定的触发器使图像在中毒前后难以区分,同时具有较好的攻击效果<sup>[18]</sup>。

### 1.2 后门防御方法

为了抵御来自后门攻击的威胁,研究者们也提出了众多防御方法。这些方法由攻击者不同的权限导致防御进入的时间点不同,可以分为两类:第一类方法针对攻击方只拥有控制数据集的权限,但后续的学习由防御方控制的场景。此场景中防御方可通过模型学习之前过滤训练集中潜在的后本样本<sup>[16]</sup>;或在模型学习的同时,基于特殊的训练策略,消除或减弱后门样

本对于模型的影响<sup>[15]</sup>。第二类是针对攻击方同时拥有数据集和模型学习控制的权限的情况。防御方只能面对一个潜在的后门模型开展防御,可试图通过移除模型中的隐藏后门以修复模型或只给出模型诊断的结论<sup>[19]</sup>;也可以在测试阶段,事先过滤可能的测试后门样本,以切断后门模型对其的响应<sup>[20]</sup>。

当攻击方仅拥有数据集的控制权限时,防御方可以通过观察输入数据的潜在特征表示来筛选异常样本。Tran 等人<sup>[21]</sup>观察到后门样本会在特征表示的协方差频谱中留下异常,他们通过其特征表示协方差矩阵的奇异值分解来过滤后门样本。Chan 等人<sup>[22]</sup>充分利用了后门样本和正常样本输入梯度的主成分分布差异,有效地将后门样本从目标类中过滤。Chen 等人<sup>[16]</sup>提出了一种基于激活聚类(Activation Clustering, AC)的方法来检测数据集集中的后门样本,通过对每个假定的目标类使用主成分分析法(PCA)<sup>[23]</sup>,获得类内样本特征层激活的降维表示,并使用 k-means ( $k=2$ )<sup>[24]</sup>进行聚类计算轮廓系数(Silhouette Score)以区分目标类。防御方也可以通过在训练过程中采用特殊的训练策略,有效地减弱注入后门的强度。Levine 等人<sup>[25]</sup>将训练集划分为多个不交叠的子集来训练多个基分类器,利用多数投票机制将模型聚合,使得后门样本不会显著地影响多数投票的结果。Hong 等人<sup>[26]</sup>发现利用基于差分隐私(Differential Privacy)的优化机制可以减弱后门。DP-SGD<sup>[15]</sup>通过在训练过程中裁剪噪声梯度,成功地防御了后门攻击,但当攻击者采用较强的后门模式时,该策略的有效性将会降低。RE 方法<sup>[27]</sup>提出了基于优化的逆向工程防御,首先通过逆向工程恢复触发器,进而准确识别目标类,并根据生成的触发器检测后门样本。

如攻击方不仅拥有控制数据集的权限还能控制模型的学习过程,防御方则可以通过将疑似的后门模型进行重构以移除后门。Liu 等人<sup>[19]</sup>观察到后门模型中某些特定的神经元对于后门样本有强烈的响应,于是将这些与后门相关的神经元进行修剪以去除模型中的

后门。Wang 等人<sup>[28]</sup>提出了一种基于合成触发器来消除隐藏后门的方法,通过异常检测器与反向工程合成后门触发器,用绝对中值差(Median Absolute Deviation)来计算合成触发器 L1 范数的异常值,基于合成的触发器对模型做剪枝和再训练,以对模型中的后门进行修补。但这些方法并不能从根源上去除后门,甚至会较大程度地影响模型对正常样本的识别性能,且所需的计算成本也非常高昂。防御方也可以通过在测试阶段过滤后门样本,以避免激活模型中的后门。如 Subedar 等人<sup>[29]</sup>利用模型不确定性来区分待测试的良性样本和后门样本。Gao 等人<sup>[20]</sup>通过输入叠加各不相同的后门触发以观察预测结果的随机性来过滤携带后门触发的测试样本。

## 2 预训练模型辅助的后门样本自过滤方法

### 2.1 方法概述

图1展示了该方法的框架,包括目标类检测和后门样本自过滤两个部分。在目标类检测阶段,首先利用预训练模型提取样本的最后一层隐藏层的输出作为特征并通过 k 近邻算法(k-nearest neighbor, kNN)<sup>[30]</sup>预测样本类别,显然预测结果与提供的样本标签并非总是一致,据此可以计算每一类样本的分类错误率,若最高错误率低于阈值,认为此数据集为良性数据集,后续完成常规模型学习,反之认为错误率最高的类别是目标类,进行后续的过滤处理。在后门样本自过滤阶段,首先使用目标类之外的  $N-1$  类样本训练部分分类网络  $F_{N-1}$ ,根据  $F_{N-1}$  对样本的分类置信度熵值从目标类中第一次过滤后门样本,由于第一次过滤结果不够精确,所以需要完整数据集进行后续过滤,即依赖模型自身的分类能力鉴别后门样本并过滤,同时将目标类中的正常样本保留,以训练完整分类模型,具体方法如下:以第一次过滤后的数据集学习完整网络,之后每训练  $K$  个轮次对模型进行一次更新,然后根据当前模型自身的分类能力,对完整数据集的目标类样本进行一轮预测,如果预测结果与其标签不一致,则判别为后

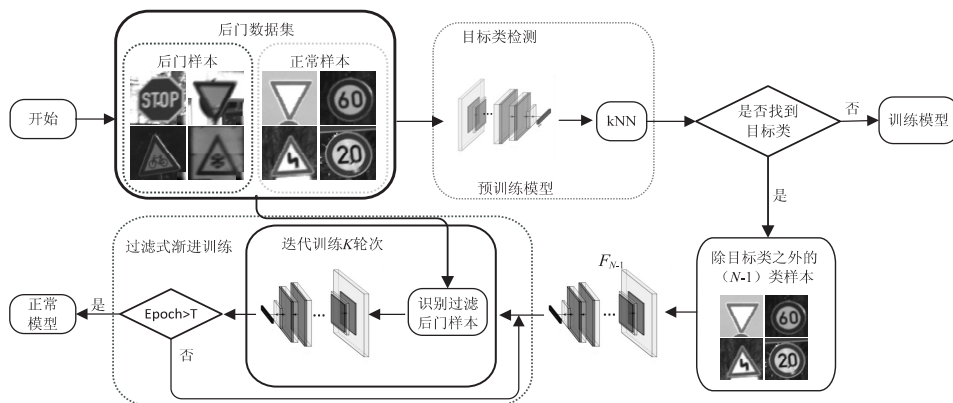


图1 方法框架



门样本并从数据集中删除,当对目标类样本完成一轮预测后,用当前过滤后的数据集继续训练模型。后门样本自过滤阶段,“后门样本过滤”和“模型学习”以交替方式多次计算,当计算次数达到指定  $T$  时,结束训练,最终在过滤后门样本的同时也得到良性模型。

## 2.2 利用预训练模型和 kNN 的目标类检测

目标类的检测方案基于以下依据:后门数据集的目标类样本由正常样本与后门样本组成,其中后门样本由来自其他类别的正常样本添加后门模式并修改标签得到。对于后门模式为局部触发器的攻击或是隐蔽性较好的触发器不可见攻击,由正常模型提取后门样本的特征表示应与其源类(后门样本中毒前的真实类别)样本的特征更相似,而与目标类别样本的特征呈现较大差异。预训练网络拥有强大特征提取能力,能够提取样本丰富的高维特征,该文采用基于对比学习的自监督训练框架得到的 ResNet-50 模型<sup>[31]</sup>作为预训练模型(下文统称为 PTM)。为了更进一步优化和凸显样本在特征空间中的分布,通过非线性数据降维 t-SNE 方法<sup>[32]</sup>,将 PTM 提取得到的高维特征降维至 2 维空间中。如上分析,目标类中的样本必定会在 2 维空间中展现出比其他正常类样本更分散的类内特征分布。然而仅通过类内的特征分散程度并不能稳定检测目标类别,因此,使用 kNN 算法检测目标类别。

图 2 展示了包含 4 个类别的部分 ImageNet 数据集经过 PTM 提取特征和 t-SNE 算法降维可视化后所

有样本的 2 维空间特征分布示意图与 kNN 分类过程。图 2(a)中攻击方式为 BadNets<sup>[11]</sup>,不同符号代表不同类别的样本特征,其中 class0 代表中毒的目标类,其他为正常类, class0 (class1)、class0 (class2)、class0 (class3) 分别表示源类为 class1、class2、class3 的后门样本。分布示意图显示,正常类别的样本特征分布比较集中,而目标类的样本分布则更分散,特别是目标类中的后门样本均分布于其源类周围,该结果与上文分析一致。

kNN 算法通过以某样本在特征空间中最相邻的  $k$  个样本中(此处  $k=5$ )多数样本的所属类别作为该样本的分类结果。因此,基于“后门样本在特征空间中近邻样本多为其源类样本”可知,用 kNN 算法预测后门样本的类别大概率是其源类,但该后门样本的给定标签为攻击者指定的目标类,如图 2(b)“分类错误”,待测样本的给定标签为 class0,其多数近邻样本标签为 class3,因此该样本经过 kNN 的分类结果为 class3,产生分类错误。与此相反,正常样本的多数近邻样本标签与该正常样本的给定标签一致,如图 2(b)“分类正确”,待测样本的给定标签为 class0,其多数近邻样本标签为 class0,因此该样本经过 kNN 的分类结果为 class0,分类正确。由以上分析不难看出,由于目标类混入了相当比例的后门样本,这导致 kNN 给出的分类错误率将高于非目标类的错误率,此规律即为目标类检测的依据。

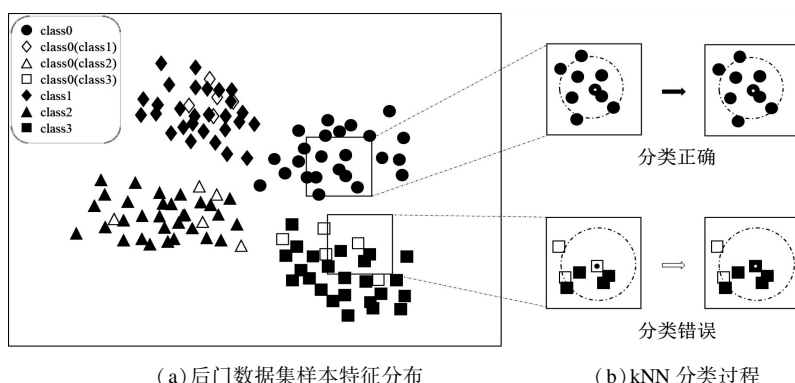


图 2 ImageNet 中存在后门样本的数据集通过 t-SNE 降维后 2 维空间特征分布示意及 kNN 分类过程

kNN 的算法步骤如下:假设数据集  $D = \{X_1, X_2, \dots, X_n\}$ ,  $X_i = (x_i, y_i)$  表示样本的图像与对应标签,其中  $y_i \in Y$ ,  $Y = \{c_1, c_2, \dots, c_m\}$  为数据集的类别集合。首先,使用 PTM 提取样本特征,以  $F(x_i)$  表示,对于待测样本  $x_i$ ,计算其与其他样本特征之间的距离,并选取前  $k$  个距离最近的样本,以出现次数最多的类别作为预测结果  $y'_i$ ,如公式(1)。

$$y'_i = \arg \max_{c_j} \sum_{x_l \in N_k(x_i)} I(y_l = c_j), \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m \quad (1)$$

式中,  $N_k(x_i)$  表示  $x_i$  最近邻的  $k$  个样本点,  $I$  为指示函

数,当  $y_i = c_i$  时,  $I$  为 1, 否则为 0, 根据多数表决规则, 决定  $x_i$  的预测类别  $y'_i$ 。在计算样本特征之间的距离时,采用欧氏距离进行度量,如公式(2),其中,  $D(x_i, x_j)$  表示  $F(x_i)$  与  $F(x_j)$  之间的距离,  $F(x_i) = \{F(x_i)^{(1)}, F(x_i)^{(2)}, \dots, F(x_i)^{(p)}\}$ ,  $p$  表示特征向量的维度,因此有:

$$D(x_i, x_j) = \left( \sum_{l=1}^p |F(x_i)^{(l)} - F(x_j)^{(l)}|^2 \right)^{\frac{1}{2}} \quad (2)$$

由于目标类中的后门样本特征上更接近源类,所以 kNN 对目标类样本的预测错误率高于正常类别,根

据这个规律,设置合适的阈值,检测后门数据集的目标类。

### 2.3 后门样本自过滤

假设数据集的类别数量为  $N$ , 检测到目标类之后, 用除目标类以外的  $N - 1$  类样本训练  $F_{N-1}$  分类网络。由于  $F_{N-1}$  没有学习目标类样本的特征, 所以对目标类的正常样本没有分类能力。不难推测, 对于目标类中的正常样本,  $F_{N-1}$  给出的所有分类节点的置信度趋彼此接近, 且都较低。相反, 后门样本均来自于其他正常类别, 且  $F_{N-1}$  对正常类别样本已具备较强的特征提取能力, 故对于后门样本,  $F_{N-1}$  将在其源类的节点上给出较高的置信度。此处借助信息熵的概念, 通过下式以表征这种区别:

$$E(x) = -\sum_{i=1}^{N-1} C_i(x) \log(C_i(x)), C_i(x) \in C(x) \quad (3)$$

式中,  $x$  表示输入样本,  $C(x)$  表示在模型分类层中所有节点的置信度的集合,  $C_i(x)$  表示第  $i$  个节点的置信度,  $E(x)$  表征了分类结果的不确定性。根据以上分析并结合信息熵的基本结论可知, 后门样本的平均  $E(x)$  将低于正常样本的平均  $E(x)$ , 从而可根据此特征确定阈值, 以实现后门样本的第一次过滤, 随后从剩余的正常样本集中, 学习初始的完整分类网络  $F_N$ 。

在第一次后门样本的过滤过程中, 不可避免会误留下一些后门样本, 同时也会错误地过滤掉部分正常样本, 从而对模型的性能产生影响。因此, 需要迭代更新模型, 以逐步增强模型对正常样本的分类性能, 同时也提升对后门样本的过滤能力。设计了如下规则: 每训练  $K$  个轮次后更新模型, 然后对目标类中的样本进行预测, 如果预测结果不是目标类, 则认为是后门样本并从数据集中删除。所有目标类样本预测结束后, 得到过滤后的数据集, 开始新的  $K$  个轮次的模型学习。这样, “后门样本过滤”和“模型学习”以交替的方式, 迭代进行  $T$  次, 得到最终的完整分类模型。从以上过程可以看出, 第一次后门样本过滤是依赖于  $F_{N-1}$  自身的分类能力, 而在后续的迭代计算中, 也是依赖于不断增强的模型自身分类能力, 逐步提升对后门样本的过滤能力。

## 3 实验结果与分析

### 3.1 实验设置

#### 3.1.1 数据集与模型

为了充分验证本方法的有效性, 分别使用 ImageNet<sup>[33]</sup>、GTSRB<sup>[34]</sup> 和 Oxford-IIIT Pets<sup>[35]</sup> 三个数据集以及端到端训练和微调两种训练方式进行实验, 它们分别包含 1 000、43 和 37 个类别。对于 ImageNet 与 GTSRB 数据集, 实验中随机选择 12 个和 14 个类别

进行实验, 以端到端的方式分别训练 DeseNet-121 与 ResNet-34 分类网络; 对于 Oxford-IIIT Pets 数据集, 在 ResNet-50 预训练模型(通过对比学习的自监督训练方式<sup>[31]</sup>得到)的基础上, 替换适应本下游任务的分类层后微调分类器。以上数据集均按照 8 : 2 划分为训练集和测试集。同时在 GTSRB 上模拟了小样本数据集, 从 43 类样本中每类随机选择 30 个左右样本组成训练集, 每类 100 个样本组成测试集, 并采用与 Oxford-IIIT Pets 任务相同的模型和训练方式微调分类器。实验中, 所有图像尺寸均为 224×224, 详细实验设置如表 1 所示。

#### 3.1.2 攻击设置

在本实验中, 分别对每个数据集中的部分样本用 BadNets<sup>[11]</sup> 和隐写<sup>[17]</sup> 两种方式植入后门, 图 3 展示了三组不同的后门样本, 其中 BadNets 的后门触发器为右下角图像尺寸为 7×7 的像素块, 隐写攻击则通过隐写的方式将此像素块嵌入图像。

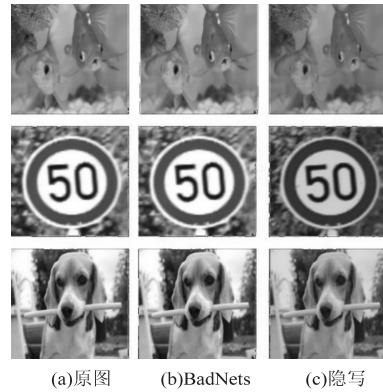


图3 样本示例

根据表 1 的设置, 使用 BadNets 和隐写攻击分别训练后门模型用于评估本方法的效果, 并在良性数据集上训练基准模型作为比较。如表 2 所示, BA (Benign Accurate) 表示对正常样本的分类准确率, ASR (Attack Success Rate) 表示攻击成功率。实验结果表明, 两种攻击方法都能在后门模型上达到较高的攻击成功率, 且对正常样本的分类性能与基准模型接近。表 2 也相应列出了相应的后门样本注入率。

#### 3.2 目标类检测性能评估

针对表 2 中的不同情况评估了本方法的性能, 并与 AC<sup>[16]</sup> 和 RE<sup>[27]</sup> 方法进行比较。本方法下, kNN 算法中的参数  $k$  设为 5, 错误率阈值设为 20%。AC 首先使用后门数据集训练得到后门网络, 用该后门网络对每一类样本提取特征并用  $k$ -means ( $k = 2$ ) 聚类方法和轮廓系数检测目标类。由于轮廓系数的值与数据集分布和模型结构密切相关, 通过实验发现, 将此阈值设为 0.45 时, 检测性能最佳。RE 方法对潜在的源-目标对进行反向工程生成触发器以检测目标类。表 3 展示

了三种方法的检测结果,⊙表示能正确检测中毒的目标类,⊗表示不能检测到目标类或检测错误。由此结果可见,AC 方法在很多情况下无法检测到目标类,RE

虽然能检测出部分 BadNets 攻击,但对隐写攻击无效,而本方法能够稳定地检测出目标类,总体表现优于 AC 和 RE。

表 1 实验设置

数据集	类别数量	训练方式	训练样本数量	测试样本数量	分类网络
ImageNet	12	end to end	12 480	3 120	DenseNet-121
GTSRB	14		20 160	5 040	ResNet-34
GTSRB(小样本)	43	fine-tuning	1 284	4 300	ResNet-50
Oxford-IIIT Pets	37		5 911	1 479	ResNet-50

表 2 基准模型和后门模型的基本指标

数据集	训练方式	基准模型 分类准确率/%	攻击方法	BA/%	ASR/%	注入率/%
ImageNet	end to end	92.24	BadNets	92.12	99.72	4.41
			隐写	89.74	68.92	4.41
GTSRB		99.96	BadNets	99.76	100	3.22
			隐写	98.67	72.78	3.22
GTSRB(小样本)	fine-tuning	95.14	BadNets	94.65	98.79	1.20
			隐写	94.12	66.88	3.21
Oxford-IIIT Pets		90.67	BadNets	90.47	99.93	2.99
			隐写	89.45	71.65	4.99

表 3 目标类检测结果

	ImageNet		GTSRB		GTSRB(小样本)		Oxford-IIIT Pets	
	BadNets	隐写	BadNets	隐写	BadNets	隐写	BadNets	隐写
AC	·	·	⊗	⊗	⊗	⊗	·	⊗
RE	·	⊗	·	⊗	·	⊗	⊗	⊗
Ours	·	·	·	·	·	·	·	·

### 3.3 后门样本自过滤性能评估

实验中,将参数  $K$  设置为 3,  $T$  设置为 15,即规定每训练 3 个轮次更新模型并过滤后门样本,以此方式迭代 15 次。表 4 使用真阳性率(true positive rate, TPR)与假阳性率(false positive rate, FPR)作为评价指标展示了最终结果,并与第一次过滤、AC 过滤结果比较,其中 TPR 表示数据集中所有正常样本被正确识别的比例,FPR 表示后门样本中被误识别为正常样本的比例。在第一次过滤时,以 2.3 节中的分类不确定度  $E(x)$  作为依据,对于 GTSRB 数据集,将阈值设为

0.02,表 4 中第一次过滤结果可以看出,GTSRB 的第一次过滤并不精确,需要进一步改善;对于 ImageNet 和 Oxford-IIIT Pets 数据集,阈值设为 1.0,在最终过滤之后,TPR 与 FPR 都得到了一定改进。AC 方法在目标类检测阶段并不能全部检测成功,这对后续的过滤是无意义的,因此,本实验假设 AC 在测试的所有情况下,都能正确检测出目标类,进而做后续的样本过滤。RE 方法的数据过滤依赖检测阶段反向生成的触发器,因此,此处只对 RE 方法检测成功的后门数据集进行后门样本过滤作为比较实验。

表 4 TPR/FPR 过滤指标比较(左边数据是 TPR,右边数据是 FPR)

	ImageNet		GTSRB		GTSRB(小样本)		Oxford-IIIT Pets	
	BadNets	隐写	BadNets	隐写	BadNets	隐写	BadNets	隐写
AC	100/0	99.96/1.27	95.44/62.77	99.99/0.15	99.21/100	99.20/75	97.21/100	97.26/97.29
RE	100/100	/	99.97/100	/	100/100	/	/	/
第一次过滤	95.31/8.18	95.31/10.73	97.44/14.15	97.44/16.46	100/33.3	99.84/50	99.15/13.56	99.20/9.83
最终结果	99.88/6.18	99.90/8.36	100/4.92	100/24.77	100/6.67	100/7.5	99.97/10.7	99.86/7.80

表4结果表明AC的过滤效果欠佳,而文中方法在不同数据集能有效地应对BadNets和隐写攻击。AC假设目标类的正常样本多于后门样本,将目标类样本按特征分布聚为两类后,认为数量较少的聚类以后门样本为主,进行删除。但在实际场景中,由于后门攻击需要一定的后门注入率及样本数量的不确定性,目标类的正常样本数量并不一定总是多于后门样本,如本实验中GTSRB小样本数据集和Oxford-IIIT Pets数据集中后门样本数量均多于目标类正常样本,因此在该场景下AC会将以正常样本为主的聚类误删除。图4(a)~(d)展示了在BadNets与隐写攻击下,AC方法对上述两种数据集目标类样本的聚类结果,其中聚类1的大部分样本为正常样本,聚类2的大部分样本为后门样本,而且聚类1样本数量少于聚类2样本数量。AC方法认为样本数量较少的聚类以后门样本为主,因此误将正常样本删除,说明不能简单通过聚类后两簇样本的相对数量关系作为过滤后门样本的依据。

RE方法的后门样本过滤依赖第一步中反向生成的后门触发器,根据结果可以看出,在此攻击设置下,RE方法并不能完美地反向生成触发器,因此FPR指标较差。

由表4数据可以看出,相比于第一次过滤,完成自过滤之后的TPR和FPR在大部分情况中都得到了进一步的改善。在GTSRB小样本数据集的BadNets和隐写攻击中,第一次过滤的TPR较为理想,但FPR高达33.3%和50%,完成自过滤之后,TPR、FPR都达到了理想水平。图5显示了GTSRB小样本数据集的两种攻击方法在迭代过程中的TPR、FPR变化趋势,图中横轴代表过滤次数,纵轴代表TPR、FPR。模型每训练3个轮次对数据集的目标类样本进行一次过滤。根据曲线可以看到,TPR一直保持在较高的理想范围,FPR则随迭代次数的增加而逐步下降,并在第3轮迭代更新后保持稳定。

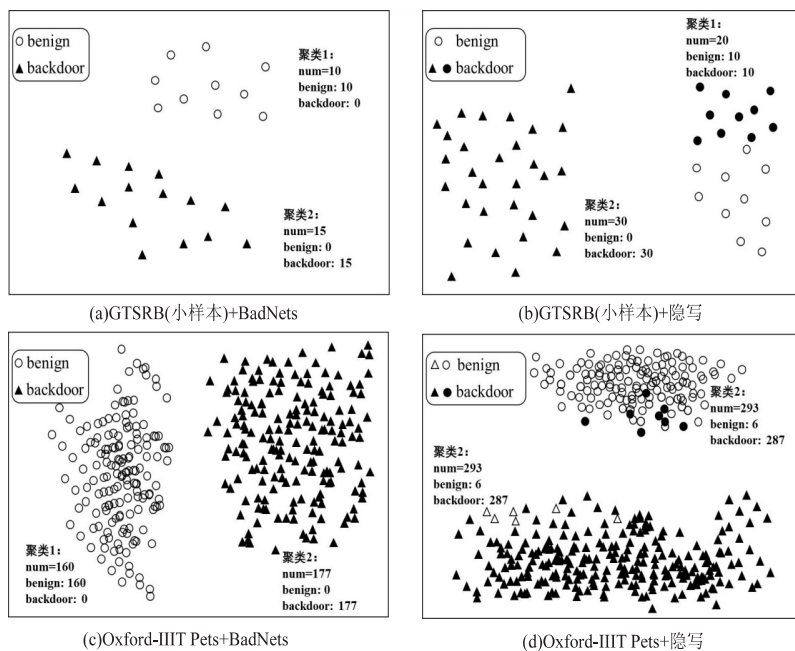


图4 在BadNets与隐写攻击方法下,AC对GTSRB(小样本)与Oxford-IIIT Pets数据集的目标类正常样本误删除示例

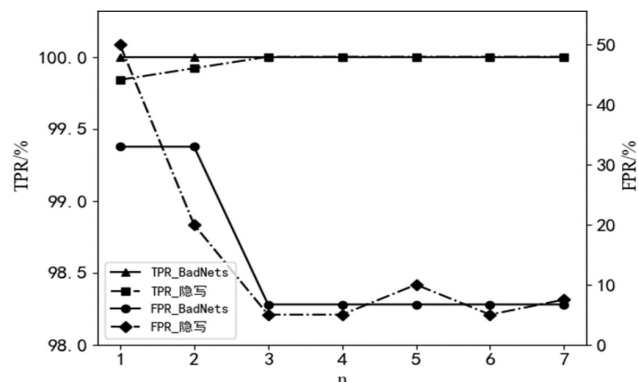


图5 注入BadNets与steganography后门样本的GTSRB小样本数据集,TPR/FPR随次数 $n$ 的变化曲线



经过后门样本自过滤的模型的 BA 与 ASR 结果如表 5 所示。大部分情况中攻击力降到了 1% 以下,同时对正常样本的分类准确率与基准模型相近。经过

后门样本自过滤的模型的 BA 与 ASR 结果如表 5 所示。大部分情况中攻击力降到了 1% 以下,同时对正常样本的分类准确率与基准模型相近。

表 5 后门样本自过滤模型分类准确率与攻击成功率

	ImageNet		GTSRB		GTSRB(小样本)		Oxford-IIIT Pets	
	BadNets	隐写	BadNets	隐写	BadNets	隐写	BadNets	隐写
BA/%	91.99	91.86	99.92	99.90	94.16	94.61	90.94	90.33
ASR/%	0.10	1.54	0.99	11.10	0.05	0.02	0.49	0.49

## 4 结束语

该文提出了一种预训练模型辅助的后门样本自过滤防御方法,实验在 ImageNet、GTSRB 和 Oxford-IIIT Pets 三个数据集和 BadNets、隐写两种攻击上进行,并与 AC 和 RE 方法做了比较。结果显示,该方法能很好地过滤后门样本并最终得到正常分类模型,在目标类检测和后门样本过滤中具有更大优势,能有效抵御后门攻击。未来工作将尝试进一步研究如何防御更复杂的后门攻击。

### 参考文献:

- [1] 卢宏涛,张秦川.深度卷积神经网络在计算机视觉中的应用研究综述[J].数据采集与处理,2016,31(1):1-17.
- [2] 张晴晴,刘勇,潘接林,等.基于卷积神经网络的连续语音识别[J].工程科学学报,2015,37(9):1212-1217.
- [3] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: learning affordance for direct perception in autonomous driving[C]//IEEE international conference on computer vision. Santiago: IEEE, 2015: 2722-2730.
- [4] 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.
- [5] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017, 30: 4768-4777.
- [6] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PloS One, 2015, 10(7): e0130140.
- [7] 田鹏,左大义,高艳春,等.面向实际场景的人工智能脆弱性分析[J].计算机技术与发展,2021,31(11):129-135.
- [8] 林加润.面向外包的云计算安全关键技术研究[D].长沙:国防科技大学,2017.
- [9] 郝建国,黄健,黄柯棣.HLA 联邦数据收集的研究与实现[J].计算机仿真,2002,19(1):38-42.
- [10] LI Y, WU B, JIANG Y, et al. Backdoor learning: a survey[J]. arXiv:2007.08745, 2020.
- [11] GU T, DOLAN-GAVITT B, GARG S. Badnets: identifying vulnerabilities in the machine learning model supply chain[J]. arXiv:1708.06733, 2017.
- [12] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv:1712.05526, 2017.
- [13] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. arXiv:1712.09665, 2017.
- [14] LIU Y, LEE W C, TAO G, et al. Abs: scanning neural networks for back-doors by artificial brain stimulation[C]//Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. London: ACM, 2019: 1265-1282.
- [15] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. Vienna: ACM, 2016: 308-318.
- [16] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv:1811.03728, 2018.
- [17] ZHU J, KAPLAN R, JOHNSON J, et al. Hidden: hiding data with deep networks[C]//Proceedings of the European conference on computer vision (ECCV). Munich: [s. n.], 2018: 657-672.
- [18] LI Y, LI Y, WU B, et al. Invisible backdoor attack with sample-specific triggers[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE, 2021: 16463-16472.
- [19] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//International symposium on research in attacks, intrusions, and defenses. [s. l.]: Springer, 2018: 273-294.
- [20] GAO Y, XU C, WANG D, et al. Strip: a defence against trojan attacks on deep neural networks[C]//Proceedings of the 35th annual computer security applications conference. San Juan: [s. n.], 2019: 113-125.
- [21] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[J]. Advances in Neural Information Processing Systems, 2018, 31: 8011-8021.
- [22] CHAN A, ONG Y S. Poison as a cure: detecting & neutralizing variable-sized backdoor attacks in deep neural networks[J]. arXiv:1911.08040, 2019.
- [23] MATEEN M, WEN J, SONG S, et al. Fundus image classification using VGG-19 architecture with PCA and SVD[J]. Symmetry, 2019, 11(1): 1.



- 
- [24] MOHAMAD I B, USMAN D. Standardization and its effects on K-means clustering algorithm [J]. Research Journal of Applied Sciences, Engineering and Technology, 2013, 6: 3299–3303.
- [25] LEVINE A, FEIZI S. Deep partition aggregation: provable defense against general poisoning attacks [J]. arXiv: 2006. 14768, 2020.
- [26] HONG S, CHANDRASEKARAN V, KAYA Y, et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping [J]. arXiv: 2002. 11497, 2020.
- [27] XIANG Z, MILLER D J, KESIDIS G. Reverse engineering imperceptible backdoor attacks on deep neural networks for detection and training set cleansing [J]. Computers & Security, 2021, 106: 102280.
- [28] 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). San Francisco: IEEE, 2019: 707–723.
- [29] SUBEDAR M, AHUJA N, KRISHNAN R, et al. Deep probabilistic models to detect data poisoning attacks [J]. arXiv: 1912. 01206, 2019.
- [30] 桑应宾. 基于 K 近邻的分类算法研究 [D]. 重庆: 重庆大学, 2009.
- [31] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//International conference on machine learning. Vienna: PMLR, 2020: 1597–1607.
- [32] LINDERMAN G C, STEINERBERGER S. Clustering with t-SNE, provably [J]. SIAM Journal on Mathematics of Data Science, 2019, 1(2): 313–332.
- [33] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database [C]//2009 IEEE conference on computer vision and pattern recognition. Miami: IEEE, 2009: 248–255.
- [34] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition [J]. Neural Networks, 2012, 32: 323–332.
- [35] ZHANG J, SHAO K, LUO X. Small sample image recognition using improved convolutional neural network [J]. Journal of Visual Communication and Image Representation, 2018, 55: 640–647.
- 
- (上接第 120 页)
- 资源管理 [J]. 计算机工程, 2021, 47(5): 169–175.
- [13] 李孜恒, 孟超. 基于深度强化学习的无线网络资源分配算法 [J]. 通信技术, 2020, 53(8): 1913–1917.
- [14] MENG F, CHEN P, WU L, et al. Power allocation in multi-user cellular networks: deep reinforcement learning approaches [J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6255–6267.
- [15] 廖晓闽, 严少虎, 石嘉, 等. 基于深度强化学习的蜂窝网资源分配算法 [J]. 通信学报, 2019, 40(2): 11–18.
- [16] LIAO X, SHI J, LI Z. A model-driven deep reinforcement learning heuristic algorithm for resource allocation in ultra-dense cellular networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69(1): 983–997.
- [17] WEI Y, YU F R, SONG M, et al. User scheduling and resource allocation in HetNets with hybrid energy supply: an actor-critic reinforcement learning approach [J]. IEEE Transactions on Wireless Communications, 2018, 17(1): 680–692.
- [18] ZHAO N, LIANG Y C, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks [J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141–5152.
- [19] FAN Y, LI H. Distributed approximating global optimality with local reinforcement learning in HetNets [C]//GLOBECOM 2017–2017 IEEE global communications conference. Singapore: IEEE, 2017: 1–7.
- [20] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C]//International conference on machine learning. New York: ICML, 2016: 1995–2003.