软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007688] [CSTR: 32375.14.jos.007688]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于有限数据的高效神经网络黑盒后门检测*

李尚松, 裴文希, 罗润洲, 李政辉, 刘晓杉, 孔维强

(大连理工大学 软件学院, 辽宁 大连 116024)

通信作者: 孔维强, E-mail: wqkong@dlut.edu.cn

摘　要: 尽管深度神经网络 (deep neural network, DNN) 已广泛应用于各个领域, 但其在敌对环境中表现出显著脆弱性. 近年来, 出现一种特殊的攻击形式, 称为后门攻击或木马攻击. 攻击者故意在 DNN 中植入后门, 带有后门的模型对正常输入做出正确的预测, 但对带有触发器的输入产生错误的预测. 为应对后门攻击的威胁, 研究人员提出各种检测方法, 然而, 这些方法依赖于严格的假设, 如白盒访问目标模型或已知触发器模式, 并且需要大量检测数据和检测时间, 这限制了在实际场景中的应用. 此外, 现有方法在后门攻击的目标类识别方面研究不足, 未能有效揭示目标类在攻击中的关键作用及其对实际任务的影响. 提出一种高效的黑盒后门检测方法 EBLD (efficient black-box backdoor detection), 在仅有少量干净数据的情况下, 结合迁移学习, 利用这些数据训练一小组良性模型和后门模型; 然后通过模型变异技术快速生成大量模型; 最后, 基于这些模型, 训练多个分类器以联合判断目标模型是否为后门模型. 在没有检测数据的情况下, 提出“特征适配”策略, 利用在其他任务上生成的良性模型和后门模型, 训练能够检测新目标模型的二元分类器. 该策略克服了对检测样本的依赖, 并且能够充分利用已有良性模型与后门模型资源, 提升检测效率. 此外, 设计基于迁移学习的二元分类器微调方法, 通过加载已有二元分类器的权重, 训练新的二元分类器, 进一步减少后门检测时间. 最后, 基于训练二元分类器时获得的最优查询集, 构建最优查询集库, 通过分析后门模型对最优查询集的输出类别分布, 快速判断后门攻击的目标类. 目标类识别不仅有助于更精确的定位攻击范围, 还能为后门防御与修复提供明确方向. 实验结果表明, 所提方法在检测精度方面显著优于现有方法, 在检测效率方面也达到较高水平.

关键词: 神经网络; 后门检测; 迁移学习; 模型变异

中图法分类号: TP309

中文引用格式: 李尚松, 裴文希, 罗润洲, 李政辉, 刘晓杉, 孔维强. 基于有限数据的高效神经网络黑盒后门检测. 软件学报. http://www.jos.org.cn/1000-9825/7688.htm

英文引用格式: Li SS, Pei WX, Luo RZ, Li ZH, Liu XS, Kong WQ. Efficient Black-box Backdoor Detection with Limited Data for Neural Networks. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7688.htm

Efficient Black-box Backdoor Detection with Limited Data for Neural Networks

LI Shang-Song, PEI Wen-Xi, LUO Run-Zhou, LI Zheng-Hui, LIU Xiao-Shan, KONG Wei-Qiang

(School of Software, Dalian University of Technology, Dalian 116024, China)

Abstract: Although deep neural networks (DNNs) have been widely used in various fields, they exhibit significant vulnerability in adversarial environments. In recent years, a specialized form of attack has emerged, called backdoor attacks or Trojan attacks. Attackers intentionally implant backdoors in DNNs, causing backdoored models to make correct predictions on normal inputs but produce incorrect predictions on inputs containing triggers. To address the threat of backdoor attacks, researchers have proposed various detection methods. However, these methods rely on strict assumptions, such as white-box access to the target model or knowledge of trigger patterns, and require a large amount of detection data and detection time, which limits their applicability in real-world scenarios. Moreover, existing methods have insufficient research on identifying target classes in backdoor attacks and fail to effectively reveal the critical role of target

* 收稿时间: 2025-06-20; 修改时间: 2026-02-26; 采用时间: 2026-04-27; jos 在线出版时间: 2026-07-08

classes in attacks and their impact on real-world tasks. This study proposes an efficient black-box backdoor detection method (EBLD). With only a small amount of clean data, transfer learning is leveraged to train a small set of benign models and backdoored models. Subsequently, a large number of models are rapidly generated through model mutation techniques. Finally, based on these models, multiple classifiers are trained to jointly determine whether the target model is a backdoored model. In the absence of detection data, this study proposes a “feature adaptation” strategy, which leverages benign models and backdoored models generated on other tasks to train binary classifiers capable of detecting new target models. This strategy overcomes the reliance on detection samples and fully utilizes existing benign and backdoored model resources, thus improving detection efficiency. In addition, a binary classifier fine-tuning method based on transfer learning is designed. By loading the weights of existing binary classifiers to train new binary classifiers, backdoor detection time can be further reduced. Finally, an optimal query set library is constructed based on the optimal query sets obtained during binary classifier training. By analyzing the output class distribution of backdoored models on the optimal query sets, the target class of the backdoor attack is quickly identified. Target class identification not only helps to more precisely locate the attack scope but also provides clear guidance for backdoor defense and repair. Experimental results demonstrate that the proposed method significantly outperforms existing methods in detection accuracy and achieves a high level of detection efficiency.

Key words: neural network; backdoor detection; transfer learning; model mutation

随着人工智能相关技术的飞速发展, 神经网络的应用日益广泛, 包括自动驾驶[1]、飞机碰撞检测[2]、无线通信[3]、军事行动[4,5]等. 在神经网络的帮助下, 人们可以解决更复杂的问题, 并取得令人满意的结果.

目前, 神经网络模型的规模越来越大, 这促使许多训练数据和计算资源不足的用户考虑将训练过程外包给第三方. 但是, 第三方提供的模型可能含有后门. 后门攻击发生在模型训练期间, 攻击者在干净样本中嵌入特殊模式并将其标签更改为错误标签, 然后将这些样本与干净样本一起作为训练集来训练后门模型. 在应用模型时, 如果输入样本中包含相应的模式, 则模型输出攻击者指定的内容[6−8], 例如, 在分类任务中, 模型产生特定的分类结果. 其中, 特殊模式称为触发器, 含有触发器的样本称为中毒样本, 中毒样本的标签称为目标标签或目标类. 后门攻击的原理如图 1 所示, 图 1 中, 定义触发器为小正方形, 位于图片的右下角, 目标类为农田类, 训练好的后门模型在推理阶段, 会将带有触发器的样本分类为目标类, 而对干净输入样本保持正确分类. 由于后门模型在正常输入下表现正常, 只有当触发器存在时才表现异常, 而检测者又很难获得与后门攻击对应的触发器, 因此, 后门攻击的检测具有挑战性. 从行业角度来看, 后门攻击是使用机器学习系统时最令人担忧的安全威胁之一[9,10].

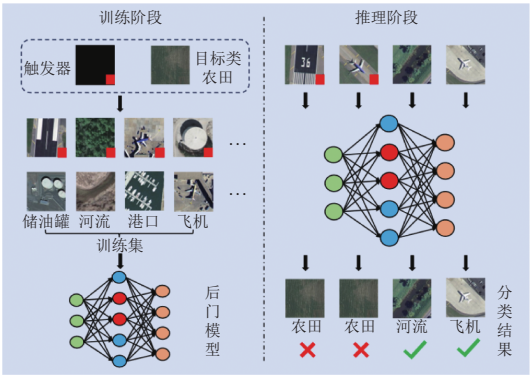


图 1 后门攻击原理

为了应对后门攻击的威胁, 已有学者提出几种检测方法[11−17], 然而这些方法在现实检测场景中存在一些局限性. 例如, Wang 等人[16]提出的 NC (neural cleanse) 方法需要白盒访问目标模型且假设较小的触发器模式. NC[16]依赖于模型梯度执行优化以生成“最小”触发器, 然后根据恢复的触发器的 L1 范数检测异常值. NC 基于一种直觉, 即在受感染的模型中, 导致错误分类到目标标签的修改要比导致错误分类到其他标签的修改小很多. NC 作为一种基于优化的方法, 使用大量的输入样本来获得良好的性能. MNTD (meta neural trojan detection)[17]方法也需要大量的输入样本, MNTD 利用大量样本训练许多良性模型和后门模型, 然后利用这些模型训练一个二元分类器来判断

目标模型是否为后门模型. Jiang 等人[18]提出一种基于图神经网络 (graph neural network, GNN) 的单类图嵌入分类 (one-class graph embedding classification, OCGEC) 框架, 用于检测模型中的后门. OCGEC 首先利用干净数据训练大量的小型模型, 然后将这些模型转换为图结构, 以利用它们的结构信息和权重. 在这些图上训练一个生成式自监督图自动编码器 (graph auto-encoder, GAE) 以学习 DNN 的表示, 并结合单类分类优化目标, 形成后门模型与正常模型间的分类边界. 该方法无需了解攻击策略, 即可检测后门模型. Wang 等人[19]提出一种基于最大边距的后门检测 (maximum-margin-based backdoor detection, MM-BD) 方法, 该方法通过分析后门攻击对分类器 Softmax 层之前的输出分布所造成的影响, 对于每个类别, 估计一个最大边距统计量. 随后, 通过对这些统计量应用无监督的异常检测器来进行检测推断. 该检测方法不需要任何检测样本, 但需要白盒访问目标模型, 此外, 该方法对于在不平衡数据上训练的干净模型会产生误判, 这是由于这类模型可能对某个类别的决策过于自信. NC 方法、MNTD 方法和 OCGEC 方法的检测效率均较低, NC 方法需要大量时间遍历模型的所有标签, 对于标签数量较多的分类模型, 检测效率大大降低. MNTD 方法和 OCGEC 方法需要消耗大量时间来训练小型模型.

本文提出一种新方法 EBLD (efficient black-box backdoor detection), 用于高效检测神经网络模型中的后门攻击. 在有限检测样本的条件下, 首先利用这些样本并结合迁移学习技术分别构建一组良性模型和后门模型, 然后, 通过模型变异技术对这些初始模型进行变异操作, 从而生成规模更大的模型集合. 最后, 基于模型集合训练分类器, 用于识别目标模型是否为后门模型. 迁移学习是机器学习中的关键技术, 其核心在于将已有知识迁移到新领域以提升学习效果[20]. 在该方法中, 迁移学习可以帮助我们在仅有少量检测数据的场景下, 训练出具有优秀性能的良性模型和后门模型, 从而有助于二元分类器的训练. 模型变异技术通过对模型参数进行微小扰动来生成新的模型, 相比于从头训练模型, 使用变异技术可以快速获得与原始模型略有不同的模型, 提高整体检测效率. 在缺乏检测样本的情况下, 我们提出了一种“特征适配”策略, 该策略的核心目标是对齐不同任务间的特征空间, 即将良性模型和后门模型的特征表示 (特征向量) 映射到与目标模型特征空间相匹配的维度. 具体而言, 例如利用 UCM 数据集[21]训练的良性模型和后门模型 (21 分类) 所生成的二元分类器, 无法直接检测基于 OPTIMAL-31 数据集[22]训练的目标模型 (31 分类), 因为两者的特征空间维度存在显著差异. 然而, 通过引入“特征适配”模块, 能够对特征表示进行映射和对齐, 从而有效利用其他任务中生成的模型来训练分类器并检测新的目标模型. 该策略不仅克服了后门检测方法对检测样本的依赖, 同时能够充分利用已有的良性模型和后门模型, 大幅提升检测效率. 此外, 为了进一步提高检测效率, 提出一种基于迁移学习的二元分类器微调方法, 该方法通过加载已有二元分类器权重, 训练新的检测目标模型的二元分类器, 从而有效减少新分类器的训练时间. 我们将目标模型对查询的输出作为其表示向量, 并将其输入二元分类器中, 以判断目标模型是否含有后门. 在分类任务中, 表示向量通常是模型对每个类别的预测概率. 为了找到最优查询集以帮助区分良性模型和后门模型, 本文引入类似于 Oh 等人[23]提出的“查询调优”技术, 生成最优查询集. 最后, 通过分析后门模型对最优查询集的输出类别分布, 进一步识别目标类. 通过识别目标类, 可以深入分析后门攻击的具体影响范围, 为模型修复提供依据. 此外, 还可设计针对目标类的防御措施, 例如对目标类相关输入进行监控或过滤触发样本, 以降低后门触发的可能性. 我们使用不同类型的后门模型对 EBLD 方法进行评估, 具体而言, 根据目标类别的不同, 包括单目标攻击后门模型和多目标攻击后门模型; 根据触发器的不同, 包括扰动攻击、补丁攻击等. 实验结果表明, EBLD 能够有效检测各类后门模型. 本文的主要贡献如下.

(1) 提出基于少量检测数据的后门攻击检测方法. 在仅有少量检测数据的情况下, 结合迁移学习训练良性模型和后门模型, 并通过模型变异技术快速生成大量模型, 从而基于这些模型训练分类器判断目标模型是否为后门模型.

(2) 提出无需检测数据的后门攻击检测方法. 在没有检测数据的情况下, 提出“特征适配”策略, 利用其他任务中生成的良性模型和后门模型, 训练二元分类器以检测新目标模型. 该策略有效避免了检测方法对检测样本的依赖, 同时能够充分利用已有的良性模型与后门模型资源.

(3) 提出后门攻击检测效率加速方法. 设计了一种基于迁移学习的二元分类器微调方法, 通过加载已有二元分类器权重, 快速训练用于检测新目标模型的分类器, 从而大幅减少检测时间.

(4) 提出后门攻击目标类判别方法. 基于训练二元分类器时获得的最优查询集, 构建最优查询集库, 通过分析后门模型对最优查询集的输出类别分布特性, 实现对后门攻击目标类的快速识别.

本文第 1 节介绍后门攻击检测方法的研究现状. 第 2 节介绍本文构建的基于有限数据的高效神经网络黑盒后门检测方法 EBLD. 第 3 节介绍实验设置. 第 4 节通过对比实验验证了 EBLD 方法的有效性. 最后总结全文. 研究总框架图如图 2 所示.

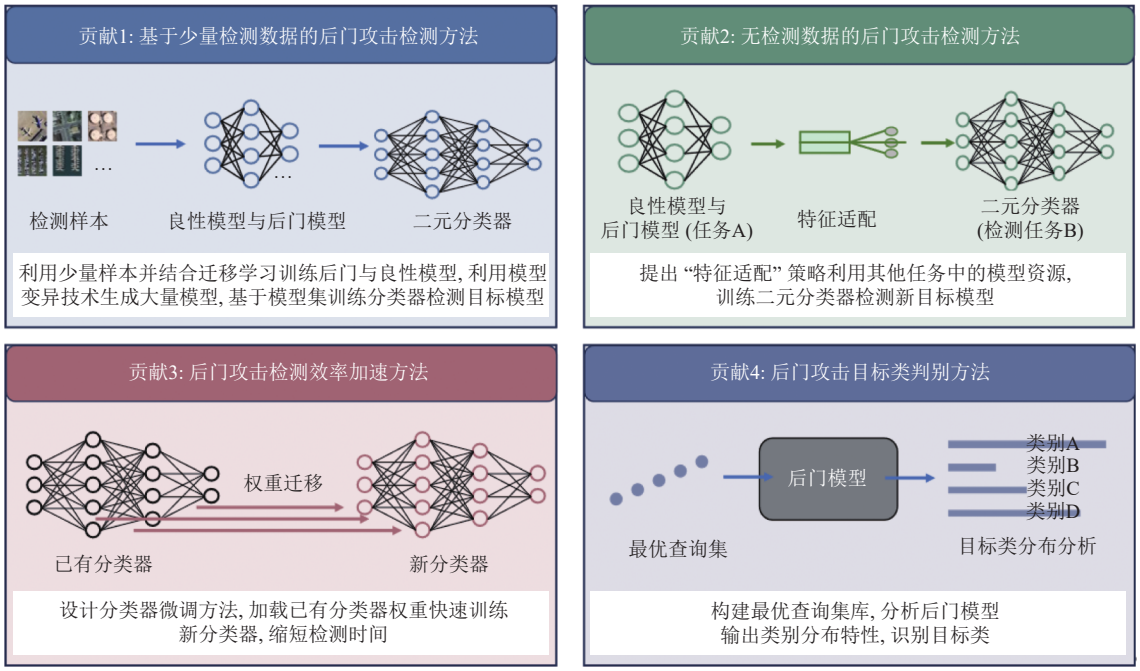


图 2　研究总框架图

1　相关工作

现有的后门检测技术包括数据集级别检测、输入级别检测和模型级别检测.

数据集级别检测的目的是检查模型的训练数据集中是否含有中毒样本. 例如, Guo 等人[13]提出一种基于聚类和质心分析的通用后门攻击防御方法, 防御的目标是通过检查训练数据集来揭示深度神经网络模型是否受到后门攻击. Al Kader Hammoud 等人[24]研究了深度神经网络在面对干净样本和中毒样本时的频率敏感性, 提出一种基于频率的中毒样本检测算法. Pan 等人[25]提出了一种基于偏移的主动分离 (active separation via offset) 检测方法, 该方法通过主动诱导模型在中毒样本与干净样本之间表现出不同的行为模式来增强两者的可分离性. 同时, 提供了一种自适应选择可疑样本数量的方法, 以实现对中毒样本的准确剔除.

输入级别检测在模型部署到操作环境之后进行, 依靠对输入样本的检查来揭示可能存在的后门. 例如, Li 等人[26]提出了非可转移性支持的后门检测技术. 该方法允许一个潜在的后门目标模型对输入数据进行分类预测, 同时利用特征提取器为输入数据和一组从预测类别中随机挑选的样本提取特征向量, 然后比较输入数据和随机样本之间的相似性, 以确定输入是触发输入还是正常输入. Fujimoto 等人[27]提出一种检测输入样本的方法, 当输入样本为正常样本时, 模型的输出概率向量会偏向防御者预设的目标标签. 而当输入样本为中毒样本时, 输出概率向量会同时偏向攻击者和防御者设定的目标标签. 他们利用这一特性来检测输入样本. Zhong 等人[14]设计了一种名为特征聚合的算法, 该算法利用干净输入, 将后门输入的特征表示分布与干净输入的特征表示分布分离. 特征聚合通过最小化干净特征表示的内部距离, 同时最大化干净输入与后门输入特征表示之间的距离来实现分离. 然后, 采用基于流的概率密度估计方法对干净输入的特征表示分布进行建模. 由于后门输入在估计分布上的可能性显著低于干净输入, 因此可以通过自适应阈值对其进行识别. Qu 等人[28]提出一种基于变形测试的顺序分析方法 (SAMT), 设

计了两种用于测试用例生成的变形关系. 通过顺序采样, 计算标签稳定率 (label stability rate, LSR), 并根据顺序概率比的变化推断输入图像是否包含触发器.

模型级别检测是指在模型部署之前, 检测模型中是否存在后门. Jiang 等人[15]提出了一种基于关键路径的后门检测方法, 该方法利用深度神经网络的可解释性检测后门攻击. Liu 等人[29]开发了一种方法来检测具有大型动态触发器的复杂后门. 该方法首先使用触发器反转技术来生成触发器, 然后检查这些触发器是否包含受害类别和目标类别之间的非自然区分特征, 该方法建立在对称特征差分技术上. Liu 等人[30]提出了一种扫描人工智能模型的技术, 该方法通过向神经元引入不同水平的刺激, 观察输出激活值如何变化来分析内部神经元的行为. 无论输入如何, 那些显著提高特定输出标签激活值的神经元被认为可能受到损害, 然后通过优化程序对触发器进行逆向工程, 以确认神经元确实受到损害. 刘叶等人[12]研究基于蒙特卡洛梯度估计的黑盒神经网络后门检测方法. Jin 等人[31]提出一种木马攻击的检测方法 CatchBackdoor, 其基于木马行为与木马路径触发错误之间的紧密联系, CatchBackdoor 从良性路径开始, 通过差分模糊测试逐步逼近木马路径, 然后, 从木马路径中逆向生成触发器, 以触发由多种木马攻击引发的错误. Zhu 等人[32]提出一种名为 NeuralSanitizer 的新方法, 用于检测和移除神经网络模型中的后门. 其识别了触发器的两个基本属性, 即在被植入后门的模型中有效、在良性模型中无效. 基于此, 采用一种新的目标函数来重构一组潜在的触发器. 然后从潜在触发器中移除冗余特征, 并利用 TNDA (transferability-based neural differential analysis) 来检测后门. 最后移除后门, 获得一个干净的模型. Zhu 等人[33]提出一种视觉-语言模型后门检测算法 (SEER), 通过在视觉和语言模态共享的特征空间中联合搜索图像触发器和恶意目标文本来检测视觉-语言模型中的后门.

尽管现有的后门检测技术取得了一定进展, 但在实际应用中仍存在局限性. 例如, 数据集级别检测通常依赖于对训练数据集的完全访问, 而购买第三方模型的用户通常无法获取训练集; 输入级别检测需要逐一检查输入样本, 其将显著降低运行效率. 现有的模型级别检测方法往往基于较强假设且检测成本较高. 为此, 本文提出一种基于有限数据的高效神经网络黑盒后门检测方法, 可同时适用于有检测数据和无检测数据的场景. 我们通过黑盒查询方式对目标模型进行模型级检测, 这种黑盒访问在现有研究中被广泛应用[17,34].

2 EBLD 后门检测

本文针对图像分类任务的神经网络模型展开研究, 目标模型的输入维度为 d_x, 输出为 C 个不同类别. 考虑后门攻击可能是单目标攻击也可能是多目标攻击, 且攻击者使用的触发器在模式、位置和大小上均可能存在显著差异.

2.1 中毒样本生成

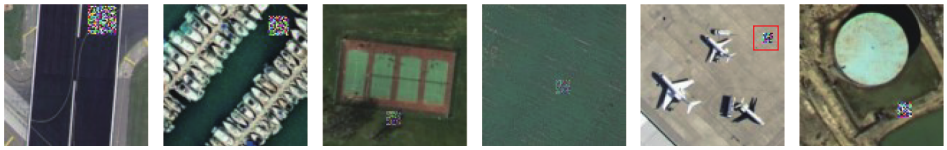
本文方法包括生成一组良性模型和后门模型, 并利用这些模型训练一个二元分类器. 为了训练后门模型, 使用通用函数 I[17]生成中毒样本, 公式如下:

$$x', y' = I(x, y; m, t, \alpha, y_t) \tag{1}$$

$$x' = (1-m)\cdot x + m\cdot((1-\alpha)t + \alpha x) \tag{2}$$

$$y' = y_t \tag{3}$$

其中, $x \in \mathbb{R}^{d_x}$ 表示干净输入, $y \in \{1,\ldots,C\}$ 表示真实标签, $m \in \{0,1\}^{d_x}$ 为触发器的掩码 (即形状和位置), $t \in \mathbb{R}^{d_x}$ 为触发器模式, α 是插入到图片的透明度, $y_t \in \{1,\ldots,C\}$ 表示中毒样本对应的恶意标签, 即目标标签. 依据函数 I 随机采样 m、t、α、y_t 可以获得不同的后门攻击设置. 具体地, 设置触发器的形状与输入图片大小相同, 或者将触发器设置为不同大小的正方形, 位置随机. 对于触发模式 t, 设置每个像素值在 [0, 1] 均匀采样. 对于透明度 α, 依据触发器的形状进行采样: 如果触发器形状与输入图片相同, α 在 [0.8, 0.95] 中均匀采样; 如果触发器形状与输入图片不同, α 以 75% 的概率从 [0.5, 0.8] 中均匀采样, 以 25% 的概率设置为 0. 图 3 展示了在 UCM 数据集上依据函数 I 生成中毒样本的部分示例, 其中, (a)–(f) 的触发器模式为随机位置的小正方形, 由方框突出显示; (h)、(j)、(l) 中, 触发器形状与原始图片相同, 混合了随机像素, 为了便于观察这种中毒样本; (g)、(i) 和 (k) 分别是与 (h)、(j) 和 (l) 相对应的原始图片.



(a) 样本1 (b) 样本2 (c) 样本3 (d) 样本4 (e) 样本5 (f) 样本6

(g) 样本7的原始图像 (h) 样本7 (i) 样本8的原始图像 (j) 样本8 (k) 样本9的原始图像 (l) 样本9

图 3 UCM 数据集生成的中毒样本示例

2.2 生成良性模型和后门模型

(1) 生成良性模型

我们采用在 ImageNet 数据集[35]上预训练的 MobileNetV2[36]作为基础模型架构训练良性模型, 预训练模型在大规模数据集上学习到丰富的特征表示, 这些特征表示可以帮助提高模型在新任务上的性能. 我们对 MobileNetV2 的分类层进行修改, 将其替换为多个全连接层, 并在部分全连接层后添加 ReLU 激活函数和 Dropout 层. 最后一个全连接层的输出维度根据不同任务的类别数进行调整. MobileNetV2 是一个典型的轻量级图像分类网络, 它通过采用线性瓶颈和倒置残差的模块设计, 在保持模型轻量和高效的同时, 有效地处理复杂特征, 在性能上取得了较好的平衡. 由于需要生成较多的良性模型和后门模型, 因此, MobileNetV2 模型的特性使其非常适合本文检测场景. 通过利用预训练的 MobileNetV2 模型, 我们获得一组性能优异的良性模型, 这些模型的输出特征能够有效支持二元分类器的训练.

针对这些良性模型, 我们进一步利用模型变异技术对每个模型进行变异. 首先创建原始模型的深拷贝模型, 包括模型的所有参数和结构. 然后遍历深拷贝模型中的参数, 为每个参数生成一个与其形状相同的随机张量, 该张量的元素服从标准正态分布, 接下来将随机张量乘以给定的扰动比例. 在本研究中, 为了引入适当的参数变化且不会导致模型性能的大幅下降, 我们将扰动幅度设置为 0.001. 最后, 将计算得到的扰动值添加到参数数据上, 从而生成变异模型. 相较于训练模型, 模型变异所需的时间更少, 可以有效扩充良性模型集和后门模型集, 提高整体后门检测效率. 算法 1 为模型变异算法.

算法 1. 生成变异模型.

输入: 原始模型 *model*, 扰动比例 *perturbation_ratio*;
输出: 变异模型 *mut_model*.

1. *mut_model* ← *copy.deepcopy*(*model*) //创建模型的副本
2. **for** *param* **in** *mut_model.parameters*() **do** //遍历新模型的所有参数
3. *random_tensor* ← *torch.randn_like*(*param*) //生成与参数形状相同的随机张量
4. *perturbation* ← *perturbation_ratio* × *random_tensor* //计算扰动值
5. *param.data.add*(*perturbation*) //将扰动值添加到参数数据上
6. **end for**
7. **return** *mut_model* //返回变异模型

算法 2 为良性模型生成算法. 该算法的输入为干净数据集以及训练良性模型 (*benign_model*) 的个数 *m*, 输出为良性模型集. 第 1 行初始化一个空列表 (*b_model*) 来存储良性模型, 第 2–8 行, 训练良性模型并生成良性变异模型, 存储到良性模型列表中, 返回良性模型集.

算法 2. 生成良性模型.

输入: 干净数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$, 良性模型个数 m;
输出: 良性模型集 *b_model*.

b_model ← [] //初始化一个空列表, 用于存储良性模型
for *model* **in** *range*(*m*) **do** //循环 *m* 次以训练 *m* 个良性模型
  benign_model ← *train_model*(*D*) //使用干净数据集 *D* 训练一个良性模型
  mutb_model ← *mut_model*(*benign_model*) //利用算法 1 对良性模型进行变异, 生成良性变异模型
  b_model.append(*benign_model*) //将训练得到的良性模型添加到模型集
  b_model.append(*mutb_model*) //将良性变异模型添加到模型集
end for
return *b_model* //返回良性模型集

(2) 生成后门模型

本文同样采用在 ImageNet 数据集上预训练的 MobileNetV2 作为基础模型架构, 并使用与训练良性模型相同的训练集进行数据投毒, 中毒率设置在 5%–50% 的范围内, 这与其他相关研究所采用的中毒率设置一致[17]. 我们根据第 2.1 节的中毒样本生成函数生成多种触发器模式, 针对每种触发器模式生成一组中毒样本, 并将这些中毒样本与干净样本混合训练一个后门模型, 从而可以得到多组后门模型. 针对这些模型, 利用算法 1 生成后门变异模型. 这些模型能够以较高的攻击成功率将带有触发器的输入样本分类为目标类, 同时在干净测试集上的准确率与良性模型相当.

算法 3 为后门模型生成算法, 输入为干净数据集、训练后门模型 (*door_model*) 的个数 *m*, 输出为后门模型集. 第 1 行初始化一个空列表 (*d_model*) 来存储后门模型, 第 3–10 行, 根据中毒率和函数 *I* 生成中毒样本并训练后门模型, 第 11–15 行, 生成后门变异模型, 将后门模型和后门变异模型存储到后门模型列表中, 返回后门模型集.

算法 3. 生成后门模型.

输入: 干净数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$, 后门模型个数 m;
输出: 后门模型集 *d_model*.

d_model ← [] //初始化一个空列表, 用于存储后门模型
for *model* **in** *range*(*m*) **do** //循环 *m* 次以训练 *m* 个后门模型
  $m,t,\alpha,y_t,p=random\_set()$ //随机采样 m、t、α、y_t 以及中毒率 p
  $indices=choose(n, int(n\times p))$ //随机选择干净数据用来生成中毒样本
  $D'\leftarrow D$ //初始化混合训练集 D'
  for *j* **in** *indices* **do** //遍历选中的干净数据
    $x'_j, y'_j=I(x_j,y_j,m,t,\alpha,y_t)$ //生成中毒样本
    $D'\leftarrow D'\cup(x'_j,y'_j)$ //将生成的中毒样本加入混合训练集 D'
  end for
  door_model ← *train_model*(D') //使用混合数据集 D' 训练后门模型
  mutd_model ← *mut_model*(*door_model*) //利用算法 1 对后门模型进行变异, 生成后门变异模型
  d_model.append(*door_model*) //将训练得到的后门模型添加到模型集
  d_model.append(*mutd_model*) //将后门变异模型添加到模型集
end for
return *d_model* //返回后门模型集

2.3 目标模型检测

2.3.1 有样本检测

在有少量干净检测样本的情况下, 我们利用这些数据训练良性模型, 并生成中毒样本训练后门模型 (第 2.1、2.2 节), 然后利用这些模型训练二元分类器以检测目标模型是否含有后门. 在模型上训练分类器需要提取模型的特征表示, 我们用 $\{(f_i,b_i)\}_{i=1}^{m}$ 表示二元分类器的训练集, 其中 f_i 代表模型, b_i 代表模型的标签, $b_i=1$ 表示模型为后门模型, $b_i=0$ 表示模型为良性模型. 我们将模型对查询集 X 的输出作为模型的特征表示, 由于在大多数输入上后门模型与良性模型的行为相似, 无法帮助区分良性模型和后门模型. 因此, 通过找到最优查询集可以在表示向量中提供最有用的信息. 具体地, 首先从高斯分布中随机抽取 k (k=10) 个初始查询; 然后利用查询调优技术[23], 即对查询集和二元分类器进行联合优化, 以尽量减少训练损失, 从而生成最优查询集 $X=\{x_1,\ldots,x_k\}$; 最后将模型的输出 $\{f_i(x_1),\ldots,f_i(x_k)\}$ 作为模型的特征表示. 我们使用两层全连接神经网络作为二元分类器的模型架构. 在检测过程中, 将最优查询集输入目标模型中, 并将其输出特征输入二元分类器进行判断. 图 4 展示了在少量检测样本条件下 EBLD 的检测过程.

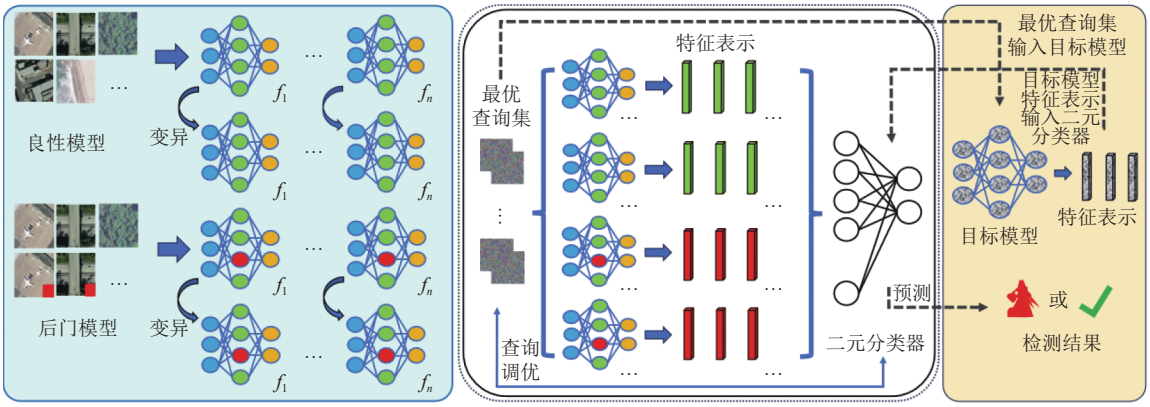


图 4 少量样本条件下 EBLD 的检测流程

2.3.2 无样本检测

在没有检测样本的情况下, 我们设计了一个“特征适配”模块, 将其他任务中生成的良性模型和后门模型的特征表示投影到目标模型的特征空间, 使其维度与目标模型的特征空间维度一致. 基于这些投影后的特征表示, 训练二元分类器来检测目标模型. “特征适配”模块的结构如表 1 所示, 其包括输入层、2 个隐藏层和 1 个输出层. 输入层接受原始特征表示, 即在其他任务中生成的良性模型和后门模型的特征表示, 维度为 input_dim. 第 1 隐藏层由 128 个神经元的全连接层构成, 通过 ReLU 激活函数增强非线性建模能力. 第 2 隐藏层包含 64 个神经元的全连接层, 同样采用 ReLU 激活函数以进一步提取特征表示. 最终, 输出层通过全连接层将隐藏空间的特征表示映射到目标特征空间, 输出特征维度为 output_dim, 对应目标任务的特征空间维度. 我们同样通过生成最优查询集来帮助区分良性模型和后门模型, 特别地, 仍然将目标模型的输出特征传入“特征适配”模块, 其主要目的是通过“特征适配”模块的作用以确保目标模型的输出分布与二元分类器训练阶段的输入分布保持一致. 这种统一的处理方式可以有效避免因分布差异导致的预测偏差, 提高检测精度. 此外, 利用其他任务中生成的良性模型和后门模型训练二元分类器, 需要确保良性模型和后门模型与目标模型的输入格式一致 (例如, 良性模型和后门模型与目标模型的输入格式同为 (3, 224, 224)), 这一要求主要是为了联合优化, 从而找到最优查询集. 然而, 本文认为在实际检测场景中, 训练与目标模型输入格式一致的良性模型和后门模型是可行且容易实现的, 该方法不仅能够在缺乏检测数据的条件下实现后门检测, 同时减少了重新训练良性模型和后门模型的时间. 为了进一步提高检测效率, 我们设计了基于迁移学习的二元分类器微调方法, 如算法 4 所示. 该方法加载检测其他目标模型的二元分类器的权重, 并将其作为新分类器的初始权重, 由于已有的二元分类器权重中蕴含了丰富的特征表达能力, 能够为新任务的学习提供良好的

基础, 因此, 可以利用少量的良性模型和后门模型对新分类器进行训练, 从而降低分类器的训练时间. 图 5 展示了无样本条件下二元分类器的训练过程.

表 1　特征适配模块结构

层次	组件	输入维度	输出维度	说明
输入	—	input_dim	—	输入特征表示, 维度为input_dim
第1隐藏层	全连接层 (Linear)	input_dim	128	将输入特征投影到128维
	激活函数 (ReLU)	128	128	引入非线性特性, 增强网络表达能力
第2隐藏层	全连接层 (Linear)	128	64	将上一层输出投影到64维
	激活函数 (ReLU)	64	64	引入非线性特性, 进一步增强网络表达能力
输出层	全连接层 (Linear)	64	output_dim	将隐藏特征映射到目标特征空间, 维度为output_dim

算法 4. 二元分类器微调算法.

输入: 已有分类器的文件路径 *saved_path_original*;
输出: 新分类器 *model*.

1. *saved_state_dict* ← *torch.load*(*saved_path_original*) //加载已有分类器的权重
2. *model_state_dict* ← *model.state_dict*() //获取新分类器的参数字典
3. **for** *name*, *param* **in** *saved_state_dict.items*() **do** //遍历已有分类器的参数
4. **if** *name* **in** *model_state_dict* and *model_state_dict*[*name*].*shape* == *param.shape* **then** //判断参数名称和形状是否匹配
5. *model_state_dict*[*name*] ← *param* //加载匹配的权重到新分类器字典
6. **end if**
7. **end for**
8. *model.load_state_dict*(*model_state_dict*, *strict* = False) //加载更新后的权重到新分类器
9. **return** *model*

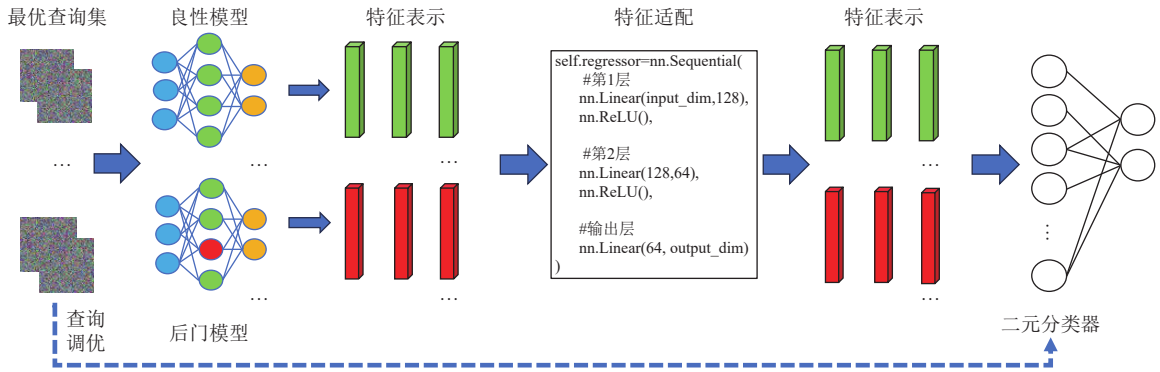


图 5　无样本条件下二元分类器的训练过程

2.3.3　联合预测机制

在后门模型检测过程中, 本文训练 5 个独立的二元分类器进行联合预测 (ensemble prediction), 旨在充分利用多个二元分类器的协作能力, 提升最终分类的准确性和稳定性. 具体而言, 这 5 个分类器分别对目标模型进行独立预测, 然后采用基于多数投票的决策机制来综合各分类器的判断结果. 如果支持后门模型的投票数超过支持良性模型的投票数, 则目标模型被分类为后门模型, 否则分类为良性模型. 投票机制表示为 $s=\sum_{i=1}^{n} y_i$, 其中, $y_i \in \{0,1\}$, $i=1,2,\ldots,5$, $y_i=1$ 表示第 i 个分类器判断目标模型为后门模型, $y_i=0$ 表示第 i 个分类器判断目标模型为良性模型,

s 表示支持后门模型的票数. 多数投票机制能够缓解单个分类器预测可能存在的偏差和不稳定性, 有效提高后门检测的稳定性. 联合预测如算法 5 所示.

算法 5. 联合预测算法.

输入: 5 个分类器的预测结果: $vote\_list = [y_1, y_2, y_3, y_4, y_5]$;
输出: 目标模型为后门模型或目标模型为良性模型.

1. $s \leftarrow 0$ //初始化后门模型投票计数
2. **for** y_i **in** *vote_list* **do** //遍历 *vote_list* 中的每个分类器预测结果
3. **if** $y_i == 1$ **then** //如果分类器预测为后门模型
4. $s \leftarrow s + 1$; //投票计数加 1
5. **end if**
6. **end for**
7. **if** $s > len(vote\_list) / 2$ **then** //如果支持后门模型的票数超过一半
8. **return** “目标模型为后门模型”
9. **else**
10. **return** “目标模型为良性模型”
11. **end if**

2.4 基于最优查询集的目标类识别机制

通过二元分类器的检测, 可以判断目标模型是否为后门模型, 如果检测结果表明目标模型为后门模型, 我们进一步识别其目标类, 以明确后门攻击的意图. 例如, 攻击者可能会选择某些任务中具有关键作用的类别作为目标类. 本文利用最优查询集识别其目标类, 如算法 6 所示. 具体而言, 将训练二元分类器过程中生成的最优查询集输入后门模型, 分析其输出的类别分布, 其中输出最集中的类别即为目标类. 为了提升目标类识别的高效性, 我们创新性地构建了一个最优查询集库, 该库预先存储针对不同输入格式的最优查询集, 可以根据后门模型的输入格式快速检索并匹配相应的最优查询集. 例如, 当后门模型的输入格式为 (3, 224, 224) 时, 在最优查询集库中直接选取格式为 (3, 224, 224) 的最优查询集用以判断后门模型的目标类. 特别地, 最优查询集库中的每组最优查询集均包含 100 个最优查询样本, 较多的查询样本能够增强统计分析的可靠性, 从而更准确地识别目标类. 这些查询样本是通过 100 个随机查询集和二元分类器联合优化获得, 在这里, 我们使用 4 层全连接神经网络作为二元分类器的模型架构. 在实际应用中, 最优查询集库的构建不仅能够显著提高目标类识别的效率与精度, 同时具有良好的通用性和可扩展性. 基于最优查询集的目标类识别机制流程如图 6 所示.

算法 6. 基于最优查询集的目标类识别算法.

输入: 后门模型 *td_model*, 最优查询集 Q;
输出: 后门模型的目标类.

1. $out \leftarrow td\_model(Q)$ //将最优查询集输入后门模型
2. $max\_values, max\_indices \leftarrow max(out)$ //获取每个样本预测分数的最大值及其对应的类别索引
3. max_list = [*index*.*item*() for *index* in *max_indices*] //将 *max_indices* 中的每个元素提取为整数并收集到 *max_list* 中
4. $index\_counts \leftarrow Counter(max\_list)$ //统计 *max_list* 中每个类别索引的出现次数
5. $most\_idx$ = None //用于存储出现次数最多的类别
6. $most\_count$ = 0 //用于存储最高的频数
7. **for** *idx*, *count* **in** *index_counts*.*items*() **do** //遍历类别及其计数

8. **if** *count* > *most_count* **then** //如果当前类别的频数大于记录的最高频数
9. *most_idx* = *idx* //更新出现次数最多的类别
10. *most_count* = *count* //更新最高的频数
11. **end if**
12. **end for**
13. *target class* ← *most_idx* //将出现频率最高的类别索引赋值为目标类
14. **return** *target class* //返回识别出的目标类

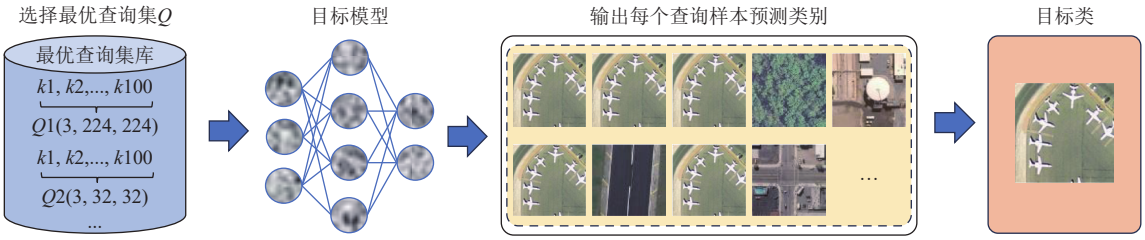


图 6 基于最优查询集的目标类识别机制流程

3 实 验

3.1 实验数据集

CIFAR10[37]数据集总共有 60 000 张彩色图像, 包括 10 个类别, 如汽车、船、马、飞机、鸟、狗等. 每个类别有 5 000 张训练集图像和 1 000 张测试集图像, 每张图像的分辨率大小为 32×32.

GTSRB[38]是德国交通标志图像数据集, 包含 43 个类别. GTSRB 数据集整体包含 51 839 张图像, 分为 39 209 张训练样本和 12 630 张测试样本.

UCM[21]数据集是从美国地质勘探局国家地图城市区域影像中心下载的. 该数据集包含 21 个土地利用场景, 如农田、飞机、海滩、建筑物、密集居民区、森林、高速公路、港口等. 每个类别包含 100 张 256×256 像素、30 cm 空间分辨率的图像.

OPTIMAL-31[22]是一个场景分类数据集, 其图像从谷歌地图收集. 该数据集包含 31 个类别, 如飞机、机场、桥梁、教堂、商业区、密集住宅区、沙漠、森林、高速公路、港口、工厂等. 每个类别由 60 张 256×256 像素的图像组成, 总共 1 860 张图像.

SIRI-WHU[39,40]谷歌影像数据集由武汉大学遥感智能数据提取与分析研究组设计, 主要覆盖中国城市地区, 该数据集包含 12 个土地利用类别, 分别为: 农场、商业区、港口、闲置用地、工业区、草地、立交桥、公园、池塘、居民区、河流和水体. 每个类别包含 200 张 200×200 像素、2 m 空间分辨率的图像.

3.2 防御设置

本节采用第 2 节中的方法, 在 5 个数据集上分别训练良性模型和后门模型. 在 UCM、OPTIMAL-31 和 SIRI-WHU 数据集中, 触发器设置为从 [16×16, 24×24, 32×32, 40×40] 中随机挑选的小正方形; 在 CIFAR10 和 GTSRB 数据集中, 触发器设置为从 [2×2, 3×3, 4×4, 5×5] 中随机挑选的小正方形. 对于每个数据集, 训练 200 个良性模型和 200 个后门模型, 并利用模型变异技术生成 200 个良性变异模型和 200 个后门变异模型. 由此, 每个数据集共包含 800 个模型, 用于训练二元分类器. 根据不同数据集样本数量, 对于 3 个遥感图像数据集, 平均每个类别标签选择 10 张左右的图片作为检测样本, 然后采用亮度增强、图片旋转等数据增强技术扩充样本数量来训练模型. 对于 CIFAR10 数据集和 GTSRB 数据集, 不进行样本数量扩充. 我们用平均准确率衡量良性模型的性能, 用平均准确率和平均攻击成功率衡量后门模型的性能, 如表 2 所示. 其中, 良性模型与后门模型均表现出较好的分类性能,

相较良性模型, 后门模型的准确率略有下降, 但攻击成功率较高, 这表明它们在一定程度上具备隐藏后门的能力.

表 2 防御模型的平均准确率和平均攻击成功率

模型	每类原始样本数	扩充样本/标签	模型属性	平均准确率 (%)	平均攻击成功率 (%)
CIFAR10-DM	100	无扩充	良性模型	45.87	—
			后门模型	42.16	74.22
GTSRB-DM	45	无扩充	良性模型	76.22	—
			后门模型	74.11	95.27
UCM-DM	8	48	良性模型	81.53	—
			后门模型	75.80	64.00
OPTIMAL-31-DM	5	39	良性模型	73.62	—
			后门模型	65.87	75.01
SIRI-WHU-DM	16	96	良性模型	83.96	—
			后门模型	79.87	77.89

注: -DM表示防御模型 (defense model)

3.3 实验设置

我们在 5 个数据集上分别训练目标模型, 以评估 EBLD 方法的性能, 并与基线方法进行比较. 我们采用在 ImageNet 数据集上预训练的 MobileNetV2 和 ResNet-50[41]作为基本模型架构, 并依据第 2.1 节中定义的中毒样本生成函数 I 生成中毒样本. 其中, 在 3 个遥感图像数据集中, 触发器掩码设置为从 [16×16, 24×24, 32×32, 40×40] 中随机挑选的小正方形, 位置随机; 在 CIFAR10 和 GTSRB 数据集中, 触发器掩码设置为从 [2×2, 3×3, 4×4, 5×5] 中随机挑选的小正方形, 位置随机; 对于触发器模式, 每个像素值从 [0, 1] 均匀采样; 透明度设置为 0. 对于每个数据集, 每个模型架构分别训练 30 个良性目标模型和 30 个后门目标模型. 表 3 展示了这些模型的平均准确率和平均攻击成功率, 其中, 后门模型的准确率与良性模型相当, 且其攻击成功率较高. 此外, 这些后门模型均为单目标攻击 (即仅有一个目标类) 模型.

表 3 目标模型的平均准确率和平均攻击成功率 (%)

模型	输入大小	基础模型架构	目标模型	平均准确率	平均攻击成功率
CIFAR10-TM	32×32×3	ResNet-50	良性模型	80.78	—
			后门模型	80.75	99.60
GTSRB-TM	32×32×3	ResNet-50	良性模型	92.45	—
			后门模型	92.67	99.79
UCM-TM	224×224×3	MobileNetV2	良性模型	85.12	—
			后门模型	80.65	76.35
		ResNet-50	良性模型	83.72	—
			后门模型	77.88	78.24
OPTIMAL-31-TM	224×224×3	MobileNetV2	良性模型	80.29	—
			后门模型	75.25	83.96
		ResNet-50	良性模型	83.68	—
			后门模型	75.34	78.95
SIRI-WHU-TM	224×224×3	MobileNetV2	良性模型	87.93	—
			后门模型	85.35	91.40
		ResNet-50	良性模型	85.38	—
			后门模型	82.14	88.60

注: -TM表示目标模型 (target model)

3.4 检测基线

我们将 EBLD 方法与 NC[16]方法、MNTD[17]方法和 MM-BD[19]方法进行对比, 其中, NC 方法和 MM-BD 方法

是白盒检测方法, 而 MNTD 方法是黑盒检测方法, NC 方法和 MNTD 方法均需要检测样本实现检测, 而 MM-BD 方法可以在不需要检测样本的情况下进行检测. 这些检测方法均为模型级别的检测. 为评估检测性能, 我们采用 ROC 曲线下的面积 (area under curve, AUC) 作为衡量指标, 这是现有后门检测方法中广泛使用的评估标准[12,17,29].

4 评　价

本节展示了在少量检测样本和无检测样本条件下, EBLD 方法的检测结果, 并将其与基线方法进行对比分析. 此外, 本节进一步评估了 EBLD 方法在泛化性方面的表现.

4.1 检测结果

4.1.1　少量检测样本条件下, EBLD 是否优于基线方法?

我们利用第 3.2 节构建的防御模型分别训练 5 个二元分类器, 记为 C-BC、G-BC、U-BC、O-BC 和 S-BC, 检测与防御模型相对应的目标模型 (第 3.3 节). 并与 NC 方法和 MNTD 方法进行比较, 检测结果如表 4 所示.

表 4　少量样本条件下 EBLD 与基线方法的 AUC 分数对比

目标模型	基础模型架构	类别数量	分类器	MNTD[17]	NC[16]	EBLD
CIFAR10-TM	ResNet-50	10	C-BC	0.7767	0.80	**0.9677**
GTSRB-TM	ResNet-50	43	G-BC	0.7733	—	**0.9511**
UCM-TM	MobileNetV2	21	U-BC	0.4784	—	**1.0**
	ResNet-50			0.7507	—	**0.9322**
OPTIMAL-31-TM	MobileNetV2	31	O-BC	0.5807	—	**1.0**
	ResNet-50			0.7216	—	**0.9022**
SIRI-WHU-TM	MobileNetV2	12	S-BC	0.6364	0.75	**1.0**
	ResNet-50			0.6311	0.83	**0.8522**

注: 加粗数据表示最优结果

从表 4 中可以看出, 针对不同的目标模型, MNTD 方法的检测 AUC 分数介于 0.45–0.80 之间. 对于 NC 方法, 我们仅检测了 CIFAR10 目标模型和 SIRI-WHU 目标模型, 这是因为, 对于类别数量较多的目标模型, NC 方法需要大量的检测时间, 以 UCM 目标模型为例, 仅检测一个 UCM 目标模型, NC 方法大约需要 1 h 53 min 19 s, 检测 60 个目标模型, NC 方法需要 113 h 19 min. 我们认为检测时间超过 72 h, 即为超时 (用“—”表示). 对于 CIFAR10 目标模型和 SIRI-WHU 目标模型, NC 方法的检测 AUC 分数介于 0.75–0.85 之间. 相比之下, EBLD 方法的检测 AUC 分数明显高于 MNTD 和 NC 方法.

4.1.2　无检测样本条件下, EBLD 是否能够有效检测?

我们结合“特征适配”策略, 利用第 3.2 节中构建的防御模型, 分别训练适应不同目标模型的二元分类器, 以在无检测样本的条件下实现后门检测. 例如, 采用基于 UCM 数据集生成的良性模型和后门模型, 训练用于检测 SIRI-WHU-TM 目标模型的二元分类器, 记为 (U-S-BC). 为验证所提方法的有效性和可靠性, 我们进行详细的检测实验, 检测结果如表 5 所示, 其中, 目标模型架构为 MobileNetV2 时记为 (M), ResNet-50 时记为 (R). 从实验结果可以看出, G-C-BC 检测 CIFAR10-TM (ResNet-50) 目标模型的 AUC 分数为 0.85, 其余所有分类器的检测 AUC 分数均大于等于 0.95, 这充分验证了 EBLD 方法在无检测样本条件下的有效性. 与 MM-BD 基线检测方法相比, 在本文方法中, U-S-BC 检测 SIRI-WHU-TM (ResNet-50) 目标模型的 AUC 分数为 0.95, 而 MM-BD 的 AUC 分数为 0.98, 略高于本文方法, 但对于其余 7 组目标模型, 本文方法的 AUC 分数均高于或等于 MM-BD 方法. 需要说明的是, MM-BD 方法通过对各类别求解最大边距统计量并使用无监督异常检测实现后门模型判定; 对于在不平衡数据上训练得到的干净模型, 其可能在某一特定类别上表现出过度自信, 从而被 MM-BD 误判. 由于 GTSRB 数据集存在显著类别不均衡, 因此, 在 GTSRB 数据集上, MM-BD 方法对目标模型的整体检测 AUC 分数仅为 0.73, 而在 SIRI-WHU 均衡的数据集上, 其检测性能 AUC 分数可达 1.0. 相比之下, EBLD 通过对模型输出分布构建二元分类器, 对

数据不平衡更具鲁棒性, 因此, 在 GTSRB 数据集上的表现更优.

表 5　无检测样本条件下 EBLD 的检测结果

方法	UCM-TM		OPTIMAL-31-TM		SIRI-WHU-TM		CIFAR10-TM	GTSRB-TM
	(M)	(R)	(M)	(R)	(M)	(R)	(R)	(R)
MM-BD[19]	0.9666	0.95	0.9666	0.93	1.0	0.98	0.8166	0.7333
C-G-BC	—	—	—	—	—	—	—	0.99
G-C-BC	—	—	—	—	—	—	0.85	—
U-O-BC	—	—	1.0	0.96	—	—	—	—
U-S-BC	—	—	—	—	1.0	0.95	—	—
O-U-BC	1.0	0.99	—	—	—	—	—	—
O-S-BC	—	—	—	—	1.0	1.0	—	—
S-U-BC	1.0	0.99	—	—	—	—	—	—
S-O-BC	—	—	1.0	1.0	—	—	—	—

我们在 5 个数据集上比较了 EBLD 方法在少量检测样本和无检测样本两种条件下的检测性能, 结果如图 7 所示. 从检测结果可以看出, 即使在无检测样本条件下, 即基于与目标模型不同的数据集所训练的良性模型和后门模型而构建的分类器, 其检测性能依然保持在较高水平. 这一结果表明, 所提出的“特征适配”策略有效解决了检测样本依赖问题, 实现了在跨任务场景下对后门攻击的高效检测.

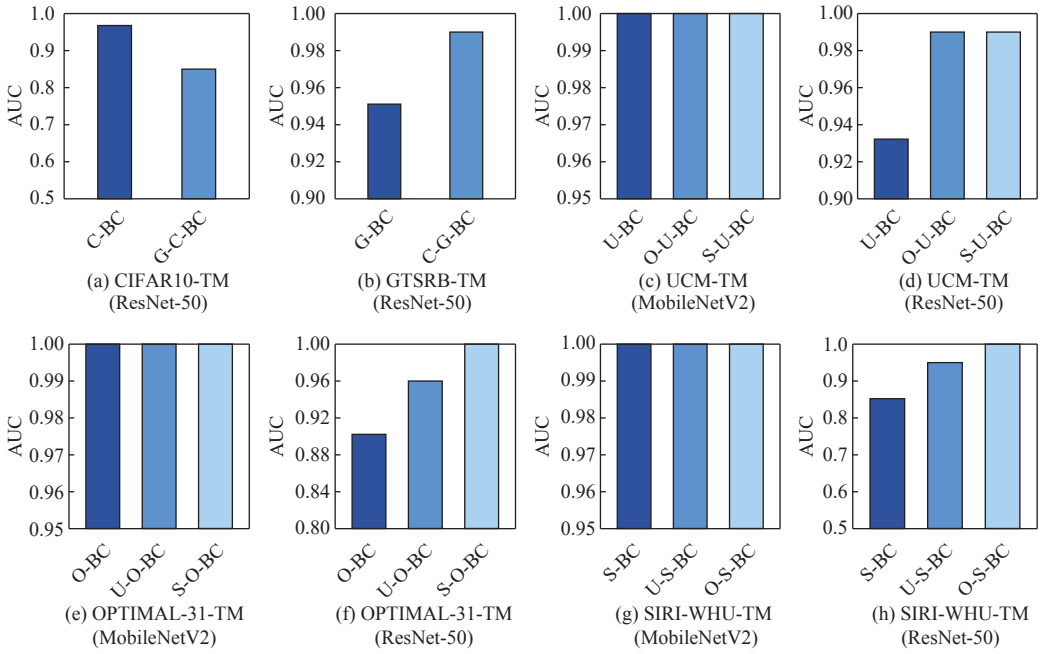


图 7　少量样本与无样本条件下检测效果的比较

为探究无检测样本条件下, EBLD 检测有效性的原因, 本文进一步对“特征适配”模块进行可解释性分析. “特征适配”模块将不同模型的输出分布映射到统一维度, 从而实现跨任务检测. 在相同的最优查询集条件下, 良性模型与后门模型会表现出深层的响应差异, 包括对查询输入的敏感性以及不同查询样本间输出模式的差异, 而不仅是面对最优查询集时表现出的置信度. 为此, 本文基于第 3.3 节在 UCM 数据集上训练的 30 个良性目标模型和 30 个后门目标模型进行实验, 计算每个模型在 10 个最优查询样本上的平均置信度, 对比结果如后文图 8 所示. 其中, 良性模型的平均置信度为 0.518, 相较于后门模型的平均置信度 0.515 略高. 因此, 仅通过置信度将无法区分良性

模型与后门模型, 说明检测器的判别依据与置信度无关, 能够确保高置信度良性模型不被误判.

4.1.3 二元分类器微调方法的有效性

本节验证了二元分类器微调方法的有效性. 具体而言, 针对目标模型 SIRI-WHU-TM 训练了两个二元分类器 U-S-BC 和 O-S-BC. 这两个分类器分别加载了 U-O-BC 和 O-U-BC 分类器的预训练权重, 并在此基础上进行再训练. 在训练阶段使用 100 个良性模型、100 个良性变异模型、100 个后门模型以及 100 个后门变异模型作为训练数据. 检测结果如图 9 所示. 实验结果表明, 通过加载预训练权重, 即使利用少量的良性模型和后门模型训练的二元分类器, 依然能够准确检测后门模型, 验证了微调方法的有效性.

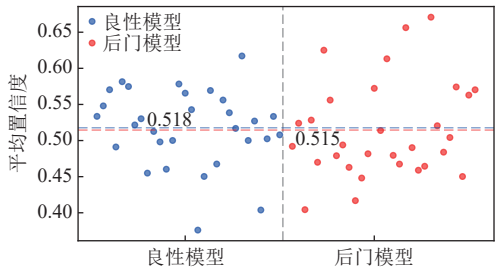


图 8 UCM 数据集上不同模型在最优查询样本上的平均置信度对比

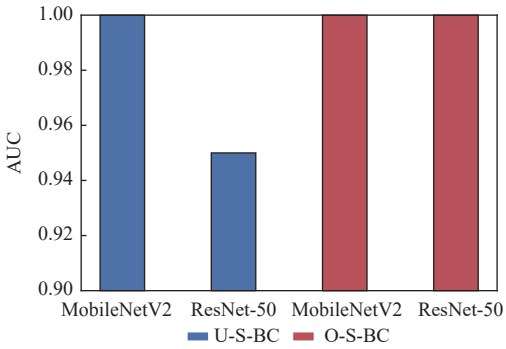


图 9 二元分类器微调方法的有效性

4.1.4 联合预测机制的优势

本节中, 我们对联合预测机制提升检测性能方面的有效性进行验证, 结果如表 6 所示. 我们采用 3 组二元分类器进行检测, 分别为 O-U-BC、U-O-BC 和 U-S-BC (每组包含 5 个分类器), 并计算每组分类器中预测的最低 AUC 分数 (用 MIN 表示) 与联合预测的 AUC 分数 (用 ENS 表示). 与最低的 AUC 分数进行比较主要是为了衡量联合预测在应对最差情况时的改进效果. 实验结果表明, 联合预测能够提升整体检测性能, 有效减小单一分类器预测误差波动, 提高检测的准确性和可靠性.

表 6 联合预测机制的有效性

模型	基础模型架构	类别	O-U-BC	U-O-BC	U-S-BC
UCM-TM	MobileNetV2	MIN	0.9977	—	—
		ENS	1.0	—	—
	ResNet-50	MIN	0.9622	—	—
		ENS	0.9966	—	—
OPTIMAL-31-TM	MobileNetV2	MIN	—	0.9977	—
		ENS	—	1.0	—
	ResNet-50	MIN	—	0.8822	—
		ENS	—	0.96	—
SIRI-WHU-TM	MobileNetV2	MIN	—	—	1.0
		ENS	—	—	1.0
	ResNet-50	MIN	—	—	0.8177
		ENS	—	—	0.95

4.1.5 基于最优查询集识别目标类的准确性评估

本节评估了基于最优查询集识别目标类的准确性. 我们采用训练 U-O-BC 分类器所得的最优查询集, 对 SIRI-WHU-TM 后门模型和 UCM-TM 后门模型进行检测, 统计识别出的目标类与真实目标类相同的概率. 结果如表 7 所示, 其中, “√”表示预测目标类正确, “×”表示预测目标类错误.

表 7 预测目标类的准确性

ID	SIRI-WHU-TM					UCM-TM				
	真实标签	预测标签 (MobileNetV2)		预测标签 (ResNet-50)		真实标签	预测标签 (MobileNetV2)		预测标签 (ResNet-50)	
		EBLD	MM-BD[19]	EBLD	MM-BD[19]		EBLD	MM-BD[19]	EBLD	MM-BD[19]
1	5	√	√	√	√	15	√	√	√	√
2	4	√	√	√	√	3	√	√	√	√
3	10	√	√	√	√	0	√	√	√	√
4	3	√	√	√	√	9	√	√	√	√
5	1	√	√	√	√	3	√	√	√	√
6	10	√	√	√	√	14	√	√	√	√
7	6	√	√	√	√	14	√	√	√	√
8	11	√	√	√	√	7	√	√	√	√
9	10	√	√	√	√	18	√	√	√	√
10	8	√	√	√	√	16	√	√	√	√
11	7	√	√	√	√	9	√	√	√	√
12	11	√	√	√	√	15	√	√	√	√
13	5	√	√	√	√	18	√	√	√	√
14	7	×	√	√	√	4	√	√	√	√
15	11	√	√	√	√	19	√	√	√	√
16	3	√	√	√	√	4	√	√	√	√
17	6	×	√	√	√	17	√	√	√	√
18	1	√	√	√	√	0	√	√	√	√
19	9	√	√	√	√	11	√	√	√	√
20	11	√	√	√	√	13	√	√	√	√
21	4	√	√	√	√	3	√	√	√	√
22	5	√	√	√	√	13	√	√	√	√
23	3	√	√	√	√	0	√	√	√	√
24	5	√	√	√	√	18	√	√	√	√
25	5	√	√	√	√	3	√	√	√	√
26	0	√	√	√	√	12	√	√	√	√
27	3	√	√	√	√	2	√	√	√	√
28	8	√	√	√	√	14	√	√	√	√
29	6	√	√	√	√	19	√	√	√	√
30	1	×	√	√	√	5	√	√	√	√
ACC (%)	—	90	100	100	100	—	100	100	100	100

从表 7 可以看出, 对于两组不同数据集的后门模型, 本文方法识别目标类的平均准确率分别为 95% 和 100%, MM-BD 方法识别目标类的平均准确率均为 100%. 因此, 在目标类识别的准确性方面, 本文方法与 MM-BD 方法的表现相当. 此外, 我们同样采用训练 U-O-BC 分类器所得的最优查询集, 计算该最优查询集识别 OPTIMAL-31-TM 后门模型目标类的准确率为 100%, 采用训练 C-BC 分类器所得的最优查询集, 计算该最优查询集识别 CIFAR10-TM 后门模型目标类的准确率为 100%, 识别 GTSRB-TM 后门模型目标类的准确率同样为 100%. 值得注意的是, 我们的最优查询集目前仅适用于识别单目标攻击中的目标类. 这是因为, 我们主要通过判断最优查询集的输出集中性进行目标类识别, 然而, 当后门模型涉及多目标攻击时, 不同目标类的输出可能呈现均匀分布的特性, 本文方法可能难以准确识别所有的目标类.

4.2 检测效率比较

本节在 UCM 目标模型上比较 EBLD 方法与基线方法的检测效率, 所有实验均在 NVIDIA RTX-A6000 GPU 上进行, 对比结果如表 8 所示. 在本文方法中, 训练一个良性模型大约需要 28 s, 训练一个后门模型大约需要 34 s, 生成一个变异模型大约需要 0.12 s, 生成良性变异模型和后门变异模型共需要约 50 s, 训练一个二元分类器大约需

要 566 s, 检测一个目标模型时间不足 1 s, 因此, EBLD 方法大约共需要 4 h 14 min 40 s. 而当我们利用已有的其他任务生成的良性模型和后门模型训练二元分类器时, 我们仅需要训练二元分类器 (EBLD-B 表示), 因此共需要 47 min 10 s. 进一步地, 通过加载已有二元分类器权重, 并利用其他任务生成的少量良性模型和后门模型进行二元分类器的再训练, 可以进一步提高检测效率, 该方案中训练一个二元分类器大约仅需要 255 s, 当训练 5 个二元分类器进行联合检测时 (记为 EBLD-F), 总耗时约 21 min 35 s. MNTD 方法大约需要 7 h 6 min 40 s. NC 方法检测一个目标模型大约需要 1 h 53 min 19 s. MM-BD 方法检测一个目标模型大约需要 3 min 12 s. 检测效率比较如表 8 所示. 此外, 本文进一步评估了 EBLD 和 EBLD-F 在不同查询数量下训练一个二元分类器的耗时表现, 如图 10 所示. 可以观察到, 随着查询次数的增加, 训练时间呈逐步上升趋势, EBLD-F 的耗时始终显著低于 EBLD. 这表明, 二元分类器的复用能有效提升检测效率.

表 8　EBLD 方法与基线方法检测效率的比较

方法	时间 (s)
NC[16]	6 799
MNTD[17]	25 600
MM-BD[19]	192
EBLD	15 280
EBLD-B	2 830
EBLD-F	1 295

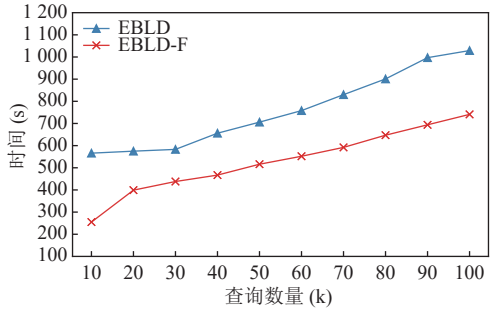


图 10　不同查询数量下 EBLD 与 EBLD-F 训练二元分类器的耗时对比

4.3　模型变异机制的参数敏感性分析

本节对模型变异机制进行参数敏感性分析. 本文在 3 个数据集上进行实验, 以已训练的防御模型为基础 (第 3.2 节), 在每个数据集内随机选取一个原始模型作为基线, 并对其施加不同程度的扰动 (0.000 5、0.001、 0.001 5、0.002), 评估变异模型的准确率以及攻击成功率. 据此考察变异模型表现随扰动幅度的变化. 实验结果如表 9 所示, 从表 9 中可以看出, 在所考察的扰动幅度范围内, 变异模型相较原始模型的准确率与攻击成功率均呈现细微差异, 说明不同扰动设置均能在保持模型稳定性的前提下, 实现有效的多样性扩充. 因此, 在合理范围内采用不同幅度的模型变异来扩充模型集是可行且稳健的, 对检测结果的整体趋势不会产生显著影响. 当扰动幅度为 0.001 时, 各数据集上的模型在稳定性与多样性之间表现出最佳平衡, 因此, 本文在实验中采用 0.001 作为默认设置.

表 9　不同扰动幅度下变异模型的性能表现 (%)

数据集	模型属性	模型性能	原始模型	变异扰动0.000 5	变异扰动0.001	变异扰动0.001 5	变异扰动0.002
GTSRB	良性模型	准确率	78.45	78.16	77.81	77.32	76.60
	后门模型	准确率	73.02	73.02	72.92	72.92	72.64
		攻击成功率	99.07	98.97	98.82	98.73	98.58
UCM	良性模型	准确率	80.95	80.24	80.24	78.57	76.90
	后门模型	准确率	77.62	77.56	76.67	73.10	68.81
		攻击成功率	72.88	72.88	69.49	64.49	64.19
SIRI-WHU	良性模型	准确率	81.67	81.67	81.67	81.25	80.00
	后门模型	准确率	81.87	81.46	81.04	79.79	78.96
		攻击成功率	88.57	85.71	81.43	80.00	78.57

4.4　不同后门模型的评估

(1) 检测不同模型架构下的适用性

为验证 EBLD 在不同模型架构下的适用性, 本文进一步在 Vision Transformer (ViT) 目标模型上进行实验. 在

UCM 数据集和 SIRI-WHU 数据集上分别训练 30 个目标模型, 每个数据集均包括 15 个良性目标模型和 15 个后门目标模型, 目标模型的准确率、攻击成功率以及检测结果如表 10 所示. 实验结果表明, EBLD 方法在 ViT 架构 (实验使用的是 ViT-tiny) 上表现出优异的检测性能. 这一结果表明, EBLD 方法不依赖于特定的模型架构, 能够有效泛化至不同类型的视觉模型, 并在不同数据集上保持稳定的检测性能.

表 10　不同数据集上 ViT 架构的模型性能与 EBLD 检测结果 (%)

数据集	良性模型	后门模型		EBLD
	准确率	准确率	攻击成功率	AUC
UCM	80.03	75.82	87.25	0.9333
SIRI-WHU	83.08	82.26	88.21	0.9466

(2) 检测不同的触发器攻击

本节评估了 EBLD 方法对不同触发器模式的检测能力, 包括扰动攻击、多触发器攻击、图像扭曲 (Wanet) 攻击、输入感知型触发器 (input-aware trigger, IaT) 攻击以及语义攻击. 对于扰动攻击, 本文在原始图片中混合了随机像素作为触发器; 对于多触发器攻击, 在原始图片上添加了两个正方形补丁. 对于语义攻击, 其触发器基于具有语义意义的目标结构, 例如在“森林”场景中添加小型水体, 并引导模型产生错误分类结果; 特别地, 由于 IaT 攻击的触发器会随输入图像变化, 表 11 展示了其一个代表性示例. 在 5 个数据集上, 共训练了 13 组后门目标模型, 其中, 在 UCM 数据集上, 采用输入感知型触发器攻击、图像扭曲攻击和语义攻击这 3 种方式, 分别训练了 10 个后门目标模型, 其余每组 30 个模型. 原始图片示例、触发模式、中毒样本示例以及检测结果如表 11 所示. 实验结果表明, EBLD 能够有效检测不同类型的触发器攻击, 表现出较高的通用性与鲁棒性.

表 11　EBLD 对不同触发器攻击的检测

后门模型	CIFAR10		GTSRB		UCM					OPTIMAL-31		SIRI-WHU	
	扰动攻击	补丁攻击	扰动攻击	补丁攻击	扰动攻击	Wanet 攻击	IaT 攻击	语义攻击	补丁攻击	扰动攻击	补丁攻击	扰动攻击	补丁攻击
原始图片													
触发模式	随机像素	多触发器	随机像素	多触发器	随机像素	图像扭曲	感知触发	语义触发	多触发器	随机像素	多触发器	随机像素	多触发器
中毒样本													
目标标签													
AUC	0.97	1.0	1.0	0.96	1.0	0.98	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(3) 检测多目标类攻击

本节评估 EBLD 方法对多目标攻击后门模型的检测能力. 与单目标攻击相比, 多目标攻击存在多个目标类别. 我们在 UCM、OPTIMAL-31 和 SIRI-WHU 数据集上, 分别生成了两组多目标攻击后门模型, 每组 30 个模型. 这些模型的触发器定义与第 3.3 节相同, 检测结果如表 12 所示, 可以看出, EBLD 方法能够有效检测多目标攻击的后门模型. 进一步地, 本文在每组多目标攻击后门模型中随机选取一个模型, 对其进行目标类识别分析. 实验结果表明, EBLD 方法能够成功识别攻击者预设目标类别中的一部分, 但尚难以覆盖完整的目标集合, 表 12 中的加粗字体 (freeway、parkinglot、beach、overpass、pond 和 industrial) 为在随机选取的多目标后门模型上识别出的目标类别. 未来工作, 我们将进一步扩展多目标识别能力, 引入多标签判别策略, 提升方法在多目标后门攻击中对全部目标类别的识别准确率.

4.5 EBLD 的适用范围与任务扩展性分析

本节分析 EBLD 方法的适用范围及其在不同任务上的扩展能力. 本文对 EBLD 方法在图像分类任务中的检

测性能进行了系统评估, 相关实验结果验证了其在多个数据集上的有效性与稳定性. 目前, EBLD 的研究范围尚未覆盖目标检测、语义分割等任务. 然而, 从方法设计角度来看, 该检测机制并不依赖于特定任务或网络结构, 其核心思想是通过二元分类器学习良性模型与后门模型在输出特征分布上的差异, 并据此判断目标模型是否含有后门, 这种基于输出分布分析的检测原理不受任务类型的限制. 在未来工作中, 我们将进一步研究 EBLD 在不同检测任务场景中的适用性与可扩展性.

表 12　EBLD 对不同目标类攻击的检测结果对比

数据集	目标类数量	目标类	AUC
UCM	5	**freeway**、golfcourse、harbor、intersection、mediumresidential	1.0
	2	overpass、**parkinglot**	1.0
OPTIMAL-31	5	airplane、airport、baseball_diamond、basketball_court、**beach**	1.0
	2	mountain、**overpass**	1.0
SIRI-WHU	5	meadow、overpass、park、**pond**、residential	1.0
	2	**industrial**、meadow	1.0

5 总　结

后门攻击对神经网络模型的安全性和可靠性构成了巨大威胁, 迫切需要开发有效的技术和工具来确保神经网络能够按预期工作, 而不受错误或漏洞的影响. 本文针对后门攻击的检测问题, 提出一种基于有限数据的高效黑盒后门检测方法 EBLD. 该方法结合迁移学习和模型变异技术, 利用少量干净数据训练良性模型和后门模型, 并通过模型变异技术快速生成大量模型, 最后利用这些模型训练多个分类器, 联合判断目标模型是否为后门模型. 在没有检测数据的情况下, 本文提出的“特征适配”策略进一步扩展了该方法的适用性, 通过利用其他任务中的良性模型和后门模型, 训练能够检测新目标模型的二元分类器, 从根本上解决了对检测样本的依赖. 为了提高检测效率, 本文设计了基于迁移学习的二元分类器微调方法, 有效减少后门检测时间. 最后, 基于训练二元分类器时获得的最优查询集, 构建最优查询集库, 并通过分析后门模型对最优查询集的输出类别分布, 实现对目标类的准确判别. 最优查询集的构建显著提升了目标类识别的效率与通用性. 目标类识别不仅为追踪攻击者提供了重要依据, 还能够指导针对性的后门防御措施设计及后门模型的修复方法. 与现有方法相比, 本文方法具有显著优势, 包括无需访问目标模型的训练集和模型参数、不依赖于对触发器性质的假设以及较高的检测效率. 在未来的研究中, 我们计划进一步扩展方法的能力, 以识别多目标攻击中的所有目标类别, 并探索其在目标检测、语义分割等任务中的适用性. 此外, 我们还计划研究后门检测策略与后门缓解策略的有效结合, 以构建更全面的防御机制, 从而进一步提升神经网络模型的安全性.

References

[1] Matloob Y, WallyAllah A, Abdel-Halim M, Sabry M, El-Tantawy S, Mostafa H. TimeFusionNet for end-to-end self-driving cars. In: Proc. of the 5th Novel Intelligent and Leading Emerging Sciences Conf. (NILES). Giza: IEEE, 2023. 42–47. [doi: 10.1109/NILES59815.2023.10296712]

[2] Zhang HH, Zhang JP, Zhong G, Liu H, Liu WQ. Multivariate combined collision detection for multi-unmanned aircraft systems. IEEE Access, 2022, 10: 103827–103839. [doi: 10.1109/ACCESS.2022.3210222]

[3] Liu D, Zhang JK, Cui JJ, Ng SX, Maunder RG, Hanzo L. Deep learning aided routing for space-air-ground integrated networks relying on real satellite, flight, and shipping data. IEEE Wireless Communications, 2022, 29(2): 177–184. [doi: 10.1109/MWC.003.2100393]

[4] Lee CE, Baek J, Son J, Ha YG. Deep AI military staff: Cooperative battlefield situation awareness for commander’s decision making. The Journal of Supercomputing, 2023, 79(6): 6040–6069. [doi: 10.1007/s11227-022-04882-w]

[5] Hodicky J, Kucuk V. Modelling and simulation and artificial intelligence for strategic political-military decision-making process: Case study. In: Proc. of the 9th Int’l Conf. on Modelling and Simulation for Autonomous Systems. Prague: Springer, 2022. 269–281. [doi: 10.1007/978-3-031-31268-7_16]

[6] Cinà AE, Grosse K, Demontis A, Biggio B, Roli F, Pelillo M. Machine learning security against data poisoning: Are we there yet?

Computer, 2024, 57(3): 26–34. [doi: 10.1109/MC.2023.3299572]

[7] TrojAI Leaderboards. 2019. https://pages.nist.gov/trojai/

[8] Lin JY, Xu L, Liu YQ, Zhang XY. Composite backdoor attack for deep neural network by mixing existing benign features. In: Proc. of the 2020 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2020. 113–131. [doi: 10.1145/3372297.3423362]

[9] IARPA. Trojans in artificial intelligence. 2019. https://www.iarpa.gov/research-programs/trojai

[10] Kumar RSS, Nyström M, Lambert J, Marshall A, Goertzel M, Comissoneru A, Swann M, Xia S. Adversarial machine learning-industry perspectives. In: Proc. of the 2020 IEEE Security and Privacy Workshops (SPW). San Francisco: IEEE, 2020. 69–75. [doi: 10.1109/SPW50608.2020.00028]

[11] Wang S, Li X, Song YL, Su M, Fu AM. A trigger sample detection scheme based on custom backdoor behaviors. Journal of Cyber Security, 2022, 7(6): 48–61. (in Chinese with English abstract) [doi: 10.19363/J.cnki.cn10-1380/tn.2022.11.03]

[12] Liu Y, Ye GH, Liu D. Backdoor detection for black-box deep neural networks based on Monte Carlo gradient estimation. Journal of Wuhan University (Natural Science Edition), 2023, 69(1): 8–18. (in Chinese with English abstract) [doi: 10.14188/j.1671-8836.2022.0071]

[13] Guo W, Tondi B, Barni M. Universal detection of backdoor attacks via density-based clustering and centroids analysis. IEEE Trans. on Information Forensics and Security, 2024, 19: 970–984. [doi: 10.1109/TIFS.2023.3329426]

[14] Zhong N, Qian ZX, Zhang XP. Backdoor attack detection via prediction trustworthiness assessment. Information Sciences, 2024, 662: 120283. [doi: 10.1016/j.ins.2024.120283]

[15] Jiang W, Wen XY, Zhan JY, Wang XP, Song ZW, Bian C. Critical path-based backdoor detection for deep neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2024, 35(3): 4032–4046. [doi: 10.1109/TNNLS.2022.3201586]

[16] Wang BL, Yao YS, Shan S, Li HY, Viswanath B, Zheng HT, Zhao BY. Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. In: Proc. of the 2019 IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2019. 707–723. [doi: 10.1109/SP.2019.00031]

[17] Xu XJ, Wang Q, Li HC, Borisov N, Gunter CA, Li B. Detecting ai trojans using meta neural analysis. In: Proc. of the 2021 IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2021. 103–120. [doi: 10.1109/SP40001.2021.00034]

[18] Jiang HY, Yu HY, Li N, Yi P. OCGEC: One-class graph embedding classification for DNN backdoor detection. In: Proc. of the 2024 Int'l Joint Conf. on Neural Networks (IJCNN). Yokohama: IEEE, 2024. 1–8. [doi: 10.1109/IJCNN60899.2024.10650468]

[19] Wang H, Xiang Z, Miller DJ, Kesidis G. MM-BD: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In: Proc. of the 2024 IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2024. 1994–2012. [doi: 10.1109/SP54263.2024.00015]

[20] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

[21] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proc. of the 18th SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. San Jose: ACM, 2010. 270–279. [doi: 10.1145/1869790.1869829]

[22] Wang Q, Liu ST, Chanussot J, Li XL. Scene classification with recurrent attention of VHR remote sensing images. IEEE Trans. on Geoscience and Remote Sensing, 2019, 57(2): 1155–1167. [doi: 10.1109/TGRS.2018.2864987]

[23] Oh SJ, Schiele B, Fritz M. Towards reverse-engineering black-box neural networks. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 121–144. [doi: 10.1007/978-3-030-28954-6_7]

[24] Al Kader Hammoud HA, Bibi A, Torr PHS, Ghanem B. Don't FREAK Out: A frequency-inspired approach to detecting backdoor poisoned samples in DNNs. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2338–2345. [doi: 10.1109/CVPRW59228.2023.00230]

[25] Pan MZ, Zeng Y, Lyu LJ, Lin X, Jia RX. ASSET: Robust backdoor data detection across a multiplicity of deep learning paradigms. In: Proc. of the 32nd USENIX Conf. on Security Symp. Anaheim: USENIX Association, 2023. 153.

[26] Li YS, Ma H, Zhang Z, Gao YS, Abuadbba A, Xue MH, Fu AM, Zheng YF, Al-Sarawi SF, Abbott D. NTD: Non-transferability enabled deep learning backdoor detection. IEEE Trans. on Information Forensics and Security, 2023, 19: 104–119. [doi: 10.1109/TIFS.2023.3312973]

[27] Fujimoto A, Yamashita S, Wang YT, Miyaji A. Backdoored-input detection by trigger embedding. In: Proc. of the 19th Asia Joint Conf. on Information Security (AsiaJCIS). Tainan, China: IEEE, 2024. 121–128. [doi: 10.1109/AsiaJCIS64263.2024.00028]

[28] Qu YB, Huang S, Chen X, Wang XY, Yao YM. Detection of backdoor attacks using targeted universal adversarial perturbations for deep neural networks. Journal of Systems and Software, 2024, 207: 111859. [doi: 10.1016/j.jss.2023.111859]

[29] Liu YQ, Shen GY, Tao GH, Wang ZT, Ma SQ, Zhang XY. Complex backdoor detection by symmetric feature differencing. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15003–15013. [doi: 10.1109/CVPR52688.2022.01458]

[30] Liu YQ, Lee WC, Tao GH, Ma SQ, Aafer Y, Zhang XY. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In: Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security. London: ACM, 2019. 1265–1282. [doi: 10.1145/3319535.3363216]

[31] Jin HB, Chen RX, Chen JY, Zheng HB, Zhang Y, Wang HH. CatchBackdoor: Backdoor detection via critical trojan neural path fuzzing. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 90–106. [doi: 10.1007/978-3-031-72970-6_6]

[32] Zhu H, Zhao Y, Zhang SZ, Chen K. NeuralSanitizer: Detecting backdoors in neural networks. IEEE Trans. on Information Forensics and Security, 2024, 19: 4970–4985. [doi: 10.1109/TIFS.2024.3390599]

[33] Zhu LW, Ning R, Li J, Xin CS, Wu HY. SEER: Backdoor detection for vision-language models through searching target text and image trigger jointly. In: Proc. of the 38h AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 7766–7774. [doi: 10.1609/aaai.v38i7.28611]

[34] Gao Y, Xu C, Wang D, Chen S, Ranasinghe DC, Nepal S. STRIP: A defence against trojan attacks on deep neural networks. In: Proc. of the 35th Annual Computer Security Applications Conf. (ACSAC 2019). San Juan: ACM, 2019. 113–125. [doi: 10.1145/3359789.3359790]

[35] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[36] Sandler M, Howard A, Zhu ML, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520. [doi: 10.1109/CVPR.2018.00474]

[37] Krizhevsky A. Learning multiple layers of features from tiny images. Toronto: University of Toronto, 2009.

[38] Stallkamp J, Schlipsing M, Salmen J, Igel C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks, 2012, 32: 323–332. [doi: 10.1016/j.neunet.2012.02.016]

[39] Zhao B, Zhong YF, Zhang LP, Huang B. The Fisher kernel coding framework for high spatial resolution scene classification. Remote Sensing, 2016, 8(2): 157. [doi: 10.3390/rs8020157]

[40] Zhu QQ, Zhong YF, Zhao B, Xia GS, Zhang LP. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geoscience and Remote Sensing Letters, 2016, 13(6): 747–751. [doi: 10.1109/LGRS.2015.2513443]

[41] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

附中文参考文献

[11] 王尚, 李昕, 宋永立, 苏铓, 付安民. 基于自定义后门的触发器样本检测方案. 信息安全学报, 2022, 7(6): 48–61. [doi: 10.19363/J.cnki.cn10-1380/tn.2022.11.03]

[12] 刘叶, 叶高含, 刘旦. 基于蒙特卡洛梯度估计的黑盒神经网络后门检测. 武汉大学学报 (理学版), 2023, 69(1): 8–18. [doi: 10.14188/j.1671-8836.2022.0071]

作者简介

李尚松, 博士生, 主要研究领域为人工智能安全, 后门检测.

裴文希, 本科生, 主要研究领域为软件工程.

罗润洲, 本科生, 主要研究领域为机器学习, 软件测试.

李政辉, 硕士, 主要研究领域为目标导向的系统验证, 深度模型黑盒测试.

刘晓杉, 博士生, 主要研究领域为金融信息化系统安全及风险, 系统质量控制.

孔维强, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为软件可靠性验证, 智能软件鲁棒性分析.