计算机科学
COMPUTER SCIENCE

联合视觉-文本特征的复合型触发器后门攻击

黄荣, 唐迎春, 周树波, 蒋学芹

引用本文

黄荣, 唐迎春, 周树波, 蒋学芹. 联合视觉-文本特征的复合型触发器后门攻击[J]. 计算机科学, 2026, 53(1): 382-394.

HUANG Rong, TANG Yingchun, ZHOU Shubo , JIANG Xueqin. Composite Trigger Backdoor Attack Combining Visual and Textual Features [J]. Computer Science, 2026, 53(1): 382-394.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200105

联合视觉-文本特征的复合型触发器后门攻击

黄　荣[1,2]　唐迎春[1]　周树波[1,2]　蒋学芹[1,2]
1 东华大学信息科学与技术学院　上海 201620
2 东华大学数字化纺织服装技术教育部工程研究中心　上海 201620

摘　要　后门攻击指攻击者通过毒化数据集,隐蔽地诱导受害模型关联中毒数据和目标标签,对人工智能技术的可信和安全产生威胁。现有后门攻击方法普遍存在着有效性和隐蔽性之间顾此失彼的矛盾,有效性强的触发器隐蔽性差,反之,隐蔽性好的触发器有效性弱。针对该问题,提出一种联合视觉-文本特征的复合型触发器净标签后门攻击。复合型触发器由通用型和个性化两部分可学习的触发器叠加而成。复合型触发器的设计和优化均以块内像素值的同余为约束,旨在诱导受害模型捕捉同余规律,建立起触发器和目标标签的关联,形成后门。通用型触发器使得中毒图像的块内像素值对位权2同余,其信号形态对于所有的中毒图像单一固定;个性化触发器使得中毒图像的边缘像素值对LoSB(Lower Significant Bit)的位权同余,其信号特定于图像的边缘位置。两部分触发器相叠加,有利于兼顾有效性和隐蔽性。在此基础上,引入CLIP(Contrastive Language-Image Pre-training)模型,联合视觉和文本特征构建驱动复合型触发器训练的监督信号。预训练的CLIP模型具有较强的泛化能力,能够引导复合型触发器吸收异类的文本特征,起到弱化图像内容特征的作用,进一步增强触发器的有效性。在CIFAR-10,ImageNet,GTSRB这3个数据集上开展了实验。结果表明,所提方法能够抵御后门防御技术的侦测,在攻击成功率指标上平均超越次优方法2.48个百分点;在峰值信噪比、结构相似性度量、梯度幅度相似性偏差和学习感知图像块相似度4项指标上分别平均超越次优方法10.61%,0.31%,68.44%和46.38%。消融实验的结果验证了联合视觉和本文特征引导复合型触发器训练的优势,还验证了通用型和个性化两部分触发器对后门攻击的有效性和隐蔽性。

关键词:后门攻击;复合型触发器;同余规律;CLIP模型
中图分类号　TP391

Composite Trigger Backdoor Attack Combining Visual and Textual Features

HUANG Rong[1,2],TANG Yingchun[1],ZHOU Shubo[1,2] and JIANG Xueqin[1,2]
1 College of Information Science and Technology,Donghua University,Shanghai 201620,China
2 Engineering Research Center of Digitized Textile & Apparel Technology,Ministry of Education,Donghua University,Shanghai 201620,China

Abstract　A backdoor attack refers to an attack covertly poisoning the dataset,subtly inducing the victim model to associate the poisoned data with a target label,thereby posing a threat to the trustworthiness and security of artificial intelligence technologies. Existing backdoor attack methods generally face a trade-off between effectiveness and stealthiness. Triggers with high effectiveness tend to lack stealthiness,while those with good stealthiness tend to have weak effectiveness. To address this issue,this paper proposes a composite trigger for clean-label backdoor attack,which combines visual and textual features. The composite trigger is composed of two learnable triggers:a universal part and an individual part. During the design and optimization of the composite trigger,pixel values within patches are constrained to follow a congruence rule. This constraint aims to induce the victim model to capture the congruence,thereby establishing an association between the trigger and the target label,forming a backdoor. The universal trigger ensures that pixel values within patches of poisoned images are congruent modulo 2,maintaining a fixed signal pattern across all poisoned images. The individual trigger,on the other hand,ensures that edge pixel values of poisoned images are congruent with respect to the weight of the LoSB,rendering its signal specific to the edge positions of each image. The two parts of the trigger are integrated to balance both effectiveness and stealthiness. Building on this,this paper introduces the CLIP model,which combines visual and textual features to construct the supervisory signal for training the composite trigger. The pre-trained CLIP model has strong generalization capabilities,enabling the composite trigger to absorb disparate textual features,which helps

到稿日期:2024-12-16　返修日期:2025-03-20
基金项目:国家自然科学基金(62001099);中央高校基本科研业务费专项资金(2232023D-30)
This work was supported by the National Natural Science Foundation of China(62001099) and Fundamental Research Fund for the Central Universities(2232023D-30).
通信作者:黄荣(rong.huang@dhu.edu.cn)

diminish the image content features and further enhances the trigger's effectiveness. Experiments are conducted on three datasets: CIFAR-10, ImageNet, and GTSRB. The results show that the proposed method can evade detection by backdoor defense techniques and outperforms the second-best method by an average of 2.48 percentage points in terms of attack success rate. Additionally, it surpasses the second-best method by an average of 10.61%, 0.31%, 68.44%, and 46.38% in peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), gradient magnitude similarity deviation (GMSD), and learned perceptual image patch similarity (LPIPS), respectively. The results of the ablation experiments demonstrate the advantage of combining visual and textual features in guiding the training of the composite trigger. These results also validate the roles of both the universal and individual triggers in enhancing the effectiveness and stealthiness of the backdoor attack.

Keywords　Backdoor attack, Composite trigger, Congruence rule, CLIP model

1　引言

近年来，由数据驱动、端到端训练的深度神经网络（Deep Neural Network，DNN）模型在自然语言处理[1-2]、语音处理[3-4]、推荐系统[5-6]和计算机视觉[7-8]等领域取得了显著突破。DNN模型的性能很大程度上取决于带标签数据集的质量和规模。为了节省成本，模型的开发者往往将数据集的收集和标注等劳务密集型工作外包给第三方机构。近些年的研究表明：攻击者伪装成数据的供应商，通过伪造少量带有触发器的恶意数据，即可在不参与模型训练的情况下，成功地施展后门攻击[9-10]（Backdoor Attack），使得受害模型在推理阶段做出有利于攻击者的决策。对后门攻击开展研究，有助于发现DNN模型的漏洞和隐患，是促使人工智能技术迈向可信和安全的重要一环。

本文聚焦于面向图像识别的后门攻击方法。典型的后门攻击过程如下。首先，攻击者设计一种具有特殊形态的信号（如棋盘像素块或加性噪声等）作为触发器，并将其与干净图像相结合，得到中毒图像；攻击者指定中毒图像的目标标签，并将其与中毒图像配组构成中毒样本。接着，攻击者伪装成数据的供应商，向模型的开发者出售带有少量中毒样本的数据集。在训练过程中，中毒样本诱导受害模型建立触发器和目标标签之间的关联，形成后门。本文将植入后门的受害模型简称为后门模型。对于干净图像，后门模型将以高准确率给出正确的预测结果。对于中毒图像，其内含的触发器将激活后门，使得后门模型的输出拟合于攻击者所指定的目标标签。

根据中毒图像和目标标签的语义是否具有一致性，可将现有后门攻击方法[11-25]划分为净标签和脏标签两大类。一般地，脏标签方法的攻击成功率较高。然而，由于图像和标签之间缺乏语义一致性，脏标签样本容易在数据清洗阶段被过滤，进而暴露攻击意图。因此，后门攻击的研究重点已逐步转向于净标签方法，主要从触发器的有效性和隐蔽性两个方面入手开展研究。触发器的有效性是指触发器在受害模型中成功植入后门的能力。触发器的隐蔽性是指触发器躲避人工检测和后门防御技术侦测的能力。通过调研发现，现有的后门攻击方法[18-25]普遍存在着触发器在有效性和隐蔽性之间顾此失彼的矛盾。有效性强的触发器隐蔽性差；反之，隐蔽性好的触发器有效性弱。

为了增强触发器的有效性，现有方法将触发器设计为单一固定模式[18-22]或视觉可见模式[18-20]。单一固定模式是指同一个触发器固定地嵌入于不同图像的相同空域位置。这意味着受害模型能够从不同中毒图像中反复学习同一个触发器的特征，从而强化触发器与目标标签之间的关联。视觉可见模式是指触发器在图像空域呈现出显著性。这意味着受害模型倾向于忽略图像的内容特征转而捕捉可见触发器的显著特征，有利于后门的植入。虽然上述方法具备了较强的触发器有效性，但隐蔽性差。对于单一固定模式而言，由于触发器在不同的中毒图像中表现出高度相似的异常行为[15]，防御者很容易侦测出触发器的存在并通过逆向工程等手段重构出触发器。对于视觉可见模式而言，显著的触发器与图像的内容之间存在差异。因而，简单的人工检测即可排查出中毒图像。

为了改善触发器的隐蔽性，现有方法[23-25]将触发器设计为样本特定模式或信息隐藏模式。样本特定模式是指对于不同中毒图像，触发器的信号形态及其嵌入的空域位置各不相同。这种特定于样本的触发器突破了后门防御技术基于“单一固定”的假设[15]，展现出较好的隐蔽性。信息隐藏模式是指将干净图像视为载体，将触发器视为隐体，利用信息隐藏技术不可见地将触发器嵌入干净图像之中，使得中毒图像与干净图像在视觉感知上保持一致，从而躲避人工检测。上述方法的触发器隐蔽性好但有效性弱。对于样本特定模式而言，受害模型在不同中毒图像中提取的触发器特征各不相同，导致触发器难以与目标标签建立起强关联。对于信息隐藏模式而言，触发器的信号强度弱、显著性差，容易被图像的内容所掩盖，使得受害模型难以捕捉触发器的特征，不利于后门的植入。

针对上述问题，本文着眼于兼顾触发器的有效性和隐蔽性，提出一种联合视觉-文本特征的复合型触发器净标签后门攻击方法。复合型触发器由通用型和个性化两部分可学习的触发器叠加而成。在优化过程中，两部分触发器均以像素值的同余为约束，旨在诱导受害模型捕捉同余规律，进而建立起触发器和目标标签的关联。通用型触发器属于单一固定模式，其信号形态为LSB（Least Significant Bit）块状噪声，即块内的像素值对位权2同余。个性化触发器特定于与图像中物体的边缘，其信号形态为LoSB（Lower Significant Bit）边缘噪声，即边缘像素值对LoSB的位权同余。由上述两部分叠加而成的复合型触发器结合了单一固定和样本特定两种模式，并利用信息隐藏实现触发器嵌入，兼顾了有效性和隐蔽性。在此基础上，本文引入CLIP（Contrastive Language-Image Pre-training）[26]模型，从联合视觉-文本多模态特征的角度，构建驱动复合型触发器训练的监督信号。基于CLIP的监督信号能够引导复合型触发器吸收异类的文本特征，起到弱化

图像内容特征的作用，有利于增强触发器的有效性。本文的主要贡献如下：

1）探讨了触发器有效性和隐蔽性之间的矛盾，结合单一固定和样本特定两种模式，提出了基于同余规律的复合型触发器，兼顾有效性和隐蔽性。

2）联合视觉-文本多模态特征的角度，引入了由宏量数据驱动预训练的 CLIP 模型，引导复合型触发器的训练，进一步增强触发器的有效性。

3）所提方法属于净标签黑盒后门攻击，具有良好的跨模型攻击能力。在 3 个常用数据集上的测试结果表明：所提方法在攻击成功率和中毒图像视觉质量方面优于现有方法，且能够抵御多种后门防御技术的侦测，具备良好的有效性和隐蔽性。

2 相关工作

2.1 后门攻击

本节以触发器的有效性和隐蔽性为线索，对现有的后门攻击方法[11-12,15-25]进行梳理，如表 1 所列。

表 1 现有后门攻击方法总结

Table 1 Overview of existing backdoor attack methods

方法	中毒样本	有效性		隐蔽性	
		单一固定	视觉可见	样本特定	信息隐藏
Gu 等[11]	脏标签	○	○	×	×
Chen 等[12]	脏标签	○	○	×	×
Turner 等[18]	净标签	○	△	×	×
Barni 等[19]	净标签	○	△	×	×
Liu 等[20]	净标签	△	△	×	×
Ning 等[21]	净标签	○	×	×	○
Zhu 等[22]	净标签	○	×	×	○
Li 等[15]	脏标签	×	×	○	○
Zhong 等[16]	脏标签	×	×	○	○
Zhang 等[17]	脏标签	×	×	○	○
Saha 等[23]	净标签	×	×	○	○
Souri 等[24]	净标签	×	×	○	○
Xu 等[25]	净标签	×	×	○	○

注：○表示属于；×表示不属于；△表示介于二者之间。

2.1.1 单一固定模式的触发器

Gu 等[11]探讨了 DNN 模型的安全性隐患，提出了一种名为 BadNets 的脏标签后门攻击方法。该方法将单一固定的棋盘像素块作为触发器，可见地粘贴于干净图像的右下角区域。Chen 等[12]基于密钥选定一幅图像作为触发器，并设置比例因子将触发器与干净图像混合。Turner 等[18]利用对抗噪声和隐特征空间插值来削弱图像的内容特征，并仿照 BadNets 将弱对比度的棋盘像素块作为触发器。Barni 等[19]在不同图像中按行叠加相同的周期性信号（如三角波、正弦波等），并通过限定周期性信号的幅值来改善中毒图像的视觉质量。Liu 等[20]利用镜面反射物理模型获得选定图像的反射模式，并将其作为触发器与干净图像叠加。Ning 等[21]建立了一种“图像到图像”的 U 型网络，旨在以可学习的方式将触发器图像浓缩为加性噪声。文献[22]将二值图像作为触发器不可见地嵌入干净图像的 LSB 位平面，获得了高视觉质量的中毒图像。

上述方法聚焦于后门攻击的有效性，触发器均属于单一固定模式。虽然许多方法通过控制叠加权重[12,21]、降低触发器信号强度[19,20]或信息隐藏技术[22]来掩盖触发器的存在性，但是触发器的信号形态未与干净图像相联系，难以抵御后门防御技术[15]的侦测。

2.1.2 样本特定模式的触发器

Li 等[15]利用自编码器网络将目标标签隐蔽地嵌入至干净图像，形成特定于干净图像的触发器。Zhong 等[16]训练 U 型网络来生成空域的微量修改概率模板，再经过多层感知器采样网络得到触发器。Zhang 等[17]提出了一种带有干扰模拟层的触发器注入网络，并以着色的边缘像素作为条件来引导中毒图像的生成。Saha 等[23]在像素空间微调中毒图像，使其在特征空间与触发器特征相接近，进而隐蔽地建立触发器与目标标签的关联。受文献[23]的启发，Souri 等[24]提出一种交替训练触发器扰动和代理模型的策略，渐进地将触发器特征与目标标签的语义特征相融合。Xu 等[25]对干净图像的 RGB 三通道施加不同的图像处理算子（如高斯模糊、直方图均衡化等），得到特定于干净图像本身的含扰图像。在此基础上，以可学习的像素级掩码为权重来结合干净图像和含扰图像，实现了扰动的隐蔽嵌入。

上述方法聚焦于后门攻击的隐蔽性，触发器的信号形态取决于干净图像，打破了不同中毒图像中触发器信号形态的一致性，能够抵御后门防御技术的侦测。然而，许多方法[15-17]依靠脏标签设置来弥补触发器有效性弱的不足，容易暴露攻击意图。

如表 1 所列，现有方法的触发器可分为单一固定和样本特定两种模式，存在着触发器有效性和隐蔽性之间顾此而失彼的矛盾。针对该问题，本文结合两种模式提出可学习的复合型触发器，并由 CLIP 引导训练，兼顾了有效性与隐蔽性。

2.2 后门防御

后门防御是后门攻击的对立面，旨在检测图像中是否存在触发器，或者移除后门模型中的后门效应（Backdoor Effect）。后门防御技术可用于评测后门攻击方法的鲁棒性和隐蔽性。

2.2.1 触发器检测

Tran 等[27]以奇异值分解为工具，提出一种基于图像特征离群值的谱签名方法来检测和删除中毒图像。STRIP（STRong Intentional Perturbation）[28]将潜在的中毒图像与干净图像叠加作为后门模型的输入，并依据输出结果的不确定程度来判断触发器的存在与否。SentiNet[29]采用 Grad-CAM（Gradient weighted Class Activation Mapping）技术[30]生成注意力热力图，并根据注意力的异常分布来检测中毒图像。进一步地，Doan 等[31]基于注意力热力图粗略地定位触发器的位置，再利用图像修复技术将触发器抹除。Chen 等[32]提出基于激活聚类的触发器检测方法，该方法对同类别图像的激活特征进行 K-means 聚类，并通过比较类簇的规模差异来判断图像中是否含有触发器。

2.2.2　后门移除

Li 等[33]提出了一种两阶段的梯度上升微调策略来破坏触发器和目标标签的关联。Zheng 等[34]按通道统计最大预激活值，并根据预激活值分布的偏态程度定位后门神经元。Liu 等[35]剪除带有异常激活值的神经元，并结合微调来进一步移除后门。Wang 等[36]提出了一种名为神经元清洗（Neural Cleanse）的后门检测方法。他们基于逆向工程依次为每一个类别构建触发器，并计算关于触发器大小的绝对中位差异常值来判决模型中是否存在后门。Zeng 等[37]提出的 I-BAU（Implicit Backdoor Adversarial Unlearning）交替地执行后门遗忘和触发器生成两项敌对的训练任务，利用所获得的隐式超梯度鲁棒地移除后门。

3　本文方法

3.1　威胁模型

本文面向基于 DNN 的图像识别模型开展后门攻击。模型的开发者致力于建立 DNN 模型 $f(\cdot;\boldsymbol{\theta})$，并由第三方提供的数据集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$ 驱动训练，学习图像域 $\boldsymbol{X}$ 到标签域 $\boldsymbol{Y}$ 的映射关系。其中 $\boldsymbol{\theta}$ 为模型的参数，$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 包含 N 幅图像，$\boldsymbol{Y}=\{1,2,\cdots,K\}$ 对应于 K 个类别的标签。

攻击者从自身非法利益出发指定欲攻击的目标类别 k。攻击者设计触发器 $\boldsymbol{\delta}$，并将其与第 k 类的干净图像 $\boldsymbol{x}$ 结合得到中毒图像，记为 $\boldsymbol{x}+\boldsymbol{\delta}$。将中毒图像和目标标签配组 $(\boldsymbol{x}+\boldsymbol{\delta},y)$ 构成中毒样本。攻击者伪装成数据的供应商，向模型的开发者出售含有中毒样本的数据集。

本文的威胁模型规定：攻击者仅能毒化一部分数据并身处黑盒攻击的场景。具体而言，攻击者无法参与受害模型 $f(\cdot;\boldsymbol{\theta})$ 的训练，并且对模型结构、损失函数以及训练细节（如批尺寸、学习率、梯度下降算法等）一无所知，仅负责制作中毒样本。在测试阶段，允许攻击者将触发器 $\boldsymbol{\delta}$ 与任意图像结合作为后门模型的输入，相应地获取后门模型的输出。但是，攻击者无法获知后门模型的内部细节，也不允许观测后门模型的推理过程。

攻击者的目标如下：

1）净标签攻击。中毒图像 $\boldsymbol{x}+\boldsymbol{\delta}$ 和目标标签 y 的语义具有一致性，即 $\boldsymbol{x}$ 属于第 k 类的图像，而 $y=k$。

2）触发器具有有效性。在训练阶段，中毒样本 $(\boldsymbol{x}+\boldsymbol{\delta},y)$ 能够诱导受害模型 $f(\cdot;\boldsymbol{\theta})$ 忽略 $\boldsymbol{x}$ 的内容特征转而将触发器 $\boldsymbol{\delta}$ 与目标标签 y 相关联，形成后门。在测试阶段，后门模型 $f(\cdot;\boldsymbol{\theta}')$ 应当将任意一幅带有触发器的图像高置信度地识别为第 k 类，即：

$$f(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{\theta}')=k \tag{1}$$

其中，$\boldsymbol{x}$ 不属于第 k 类，即 $y\neq k$；$\boldsymbol{\theta}'$ 为后门模型的参数。同时，后门模型 $f(\cdot;\boldsymbol{\theta}')$ 应当对干净图像给出正确的识别结果，即：

$$f(\boldsymbol{x};\boldsymbol{\theta}')=y \tag{2}$$

3）触发器具有隐蔽性。（1）中毒图像和干净图像在视觉感知上应当具有一致性，即 $\boldsymbol{x}+\boldsymbol{\delta}\approx\boldsymbol{x}$，以躲避人工检测。（2）与干净图像相比，中毒图像不具有特殊的异常行为[15]，能够躲避后门防御技术 g 的侦测，即 $g(\boldsymbol{x}+\boldsymbol{\delta})\approx g(\boldsymbol{x})$。

3.2　方法概述

本文提出一种联合视觉-文本特征的复合型触发器净标签后门攻击方法，整体框架如图 1 所示。首先，设计由通用型和个性化两部分触发器叠加而成的可学习的复合型触发器。在优化过程中，两部分触发器均以块内像素值的同余为约束，根据反向传播梯度的正负进行更新。通用型触发器属于单一固定模式，其信号形态为 LSB 块状噪声；个性化触发器属于样本特定模式，其信号形态为 LoSB 边缘噪声。它们叠加而成的复合型触发器结合了两类模式的优势，兼顾了触发器的有效性和隐蔽性。在此基础上，本文引入由预训练文本和图像编码器组成的 CLIP 模型，利用 CLIP 模型筛选出与干净图像相似度最低的异类文本，并构建基于视觉-文本多模态特征的对比损失函数。基于 CLIP 的损失函数驱动复合型触发器的训练，旨在将异类的文本特征融入复合型触发器之中，进而起到弱化图像内容特征的作用。此外，由于攻击者不具备受害模型的知识，需建立预训练的代理模型作为梯度反向传播的桥梁。

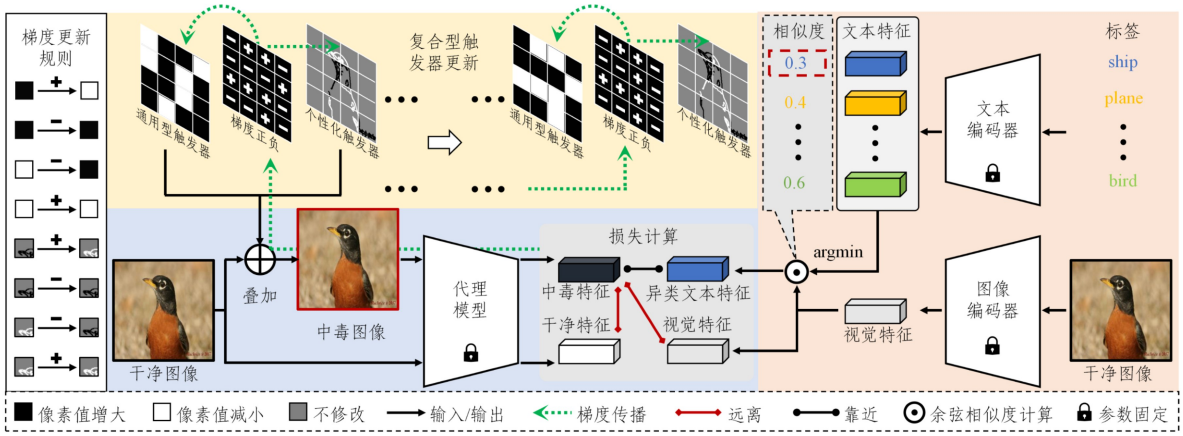


图 1　联合视觉-文本特征的复合型触发器训练框架

Fig. 1　Framework of training composite trigger based on joint visual-textual features

3.3　同余规律

同余规律是指图像块内的全部或部分像素值对某一数值的余数均相等。复合型触发器的通用型和个性化两部分均以同余规律为基础。

具体地，将图像非重叠地分割成 $S\times S$ 大小的块，记为 $\boldsymbol{p}=\{p_i\mid 1\leqslant i\leqslant S^2\}$。设二值的取样掩码为 $\boldsymbol{m}=\{m_i\mid 1\leqslant i\leqslant S^2\}$。将二进制的位权记为 2^n，对于 8 比特图像而言，$0\leqslant n\leqslant 7$。某图像块内，按掩码 $\boldsymbol{m}$ 取样的像素值对于位权 2^n 同余的规律可表达为：

$$\mathrm{mod}(p_i,2^n)\cdot\mathbf{1}(m_i=1)=\mathrm{mod}(p_j,2^n)\cdot\mathbf{1}(m_j=1) \quad (3)$$

其中，$\mathbf{1}(\cdot)$为指示函数，$1\leqslant i,j\leqslant S^2$。本文利用信息隐藏技术将触发器嵌入干净图像，使得中毒图像的像素值符合同余规律。具体操作为：

$$p_i'=\begin{cases}p_i-\mathrm{mod}(p_i,2^n)+\delta_i, & m_i=1\\ p_i, & m_i=0\end{cases} \quad (4)$$

其中，p_i'表示嵌入触发器后的像素值。

触发器满足以下条件：1)对于 $m_i=0$ 的位置，δ_i 为空(无定义)；2)对于 $m_i=1$ 的位置，δ_i 的取值为 0 到 2^n-1 之间的整数，且 $\delta_i\cdot\mathbf{1}(m_i=1)=\delta_j\cdot\mathbf{1}(m_j=1)$，$1\leqslant i,j\leqslant S^2$。这意味着在图像块内，按掩码 $\boldsymbol{m}$ 取样的 p_i'对位权 2^n 的余数均为 δ_i，符合同余规律。

3.4 复合型触发器

复合型触发器由通用型和个性化两部分可学习的触发器按 1∶1 比例逐像素相加而成，结合了单一固定和样本特定两类模式。两部分触发器均以块内像素值同余为约束进行设计和更新。

通用型触发器属于单一固定模型，其信号形态为 LSB 块状噪声。设通用型触发器中某一 $S\times S$ 大小的块为$\boldsymbol{\delta}^{\mathrm{U}}=\{\delta_i^{\mathrm{U}}\mid 1\leqslant i\leqslant S^2\}$，对应的二值取样掩码为$\boldsymbol{m}^{\mathrm{U}}=\{m_i^{\mathrm{U}}=1\mid 1\leqslant i\leqslant S^2\}$，即$\boldsymbol{m}^{\mathrm{U}}$ 固定为全 1。因此，对于通用型触发器而言，$\delta_i^{\mathrm{U}}=\delta_j^{\mathrm{U}}$ 成立，其中 $1\leqslant i,j\leqslant S^2$。通用型触发器仅修改干净图像的 LSB 位平面，$\delta_i^{\mathrm{U}}$ 的取值为 0 或 1，记为 $\delta_{\min}^{\mathrm{U}}=0$ 和 $\delta_{\max}^{\mathrm{U}}=1$。可知，嵌入通用型触发器后，同一块内的像素值同余。

个性化触发器属于样本特定模式，其信号形态为 LoSB 边缘噪声。设通用型触发器中某一 $S\times S$ 大小的块为$\boldsymbol{\delta}^{\mathrm{I}}=\{\delta_i^{\mathrm{I}}\mid 1\leqslant i\leqslant S^2\}$，对应的二值取样掩码$\boldsymbol{m}^{\mathrm{I}}=\{m_i^{\mathrm{I}}=1\mid 1\leqslant i\leqslant S^2\}$为块 $\boldsymbol{p}$ 的边缘检测结果，即：若 p_i 为边缘像素，$m_i^{\mathrm{I}}=1$；否则 $m_i^{\mathrm{I}}=0$。为避免与通用型触发器产生冲突，个性化触发器修改除 LSB 以外若干层 LoSB 位平面中的边缘像素，即 $n\geqslant 1$。因此，δ_i^{I} 的取值范围为 0 到 $2^{n+1}-2$ 之间的偶数。特别地，本文规定 δ_i^{I} 只取 0 或 $2^{n+1}-2$，记为 $\delta_{\min}^{\mathrm{I}}=0$ 和 $\delta_{\max}^{\mathrm{I}}=2^{n+1}-2$，而排除其他偶数。这增加了不同块之间 $\boldsymbol{\delta}^{\mathrm{I}}$ 的取值差异，有利于增强同余规律的显著性。就单像素而言，虽然个性化触发器的最大修改量 $\delta_{\max}^{\mathrm{I}}>\delta_{\max}^{\mathrm{U}}$，但其修改集中于图像的边缘像素。换言之，个性化触发器仅影响图像的高频分量，对视觉感知的损害有限，具有隐蔽性。

复合型触发器定义为 $\boldsymbol{\delta}=\boldsymbol{\delta}^{\mathrm{U}}+\boldsymbol{\delta}^{\mathrm{I}}$。按式(4)将复合型触发器嵌入干净图像后，块内的像素值符合下列同余规律：1)非边缘($m_i^{\mathrm{U}}=1$ 且 $m_i^{\mathrm{I}}=0$)处的像素值对位权 2 同余；2)边缘($m_i^{\mathrm{U}}=1$ 且 $m_i^{\mathrm{I}}=1$)处的像素值既对位权 2 同余，又对位权 2^{n+1}同余。数据集中所有中毒图像的像素值均以块为单位对位权 2 同余，符合相同的同余规律，属于单一固定模式。同时，由于个性化触发器的二值取样掩码 $\boldsymbol{m}^{\mathrm{I}}$ 与图像的边缘像素挂钩，因此，不同图像的个性化触发器各不相同，属于样本特定模式。可见，本文所设计的复合型触发器结合了单一固定和样本特定两类模式的优势，兼顾了触发器的有效性和隐蔽性。

将反向传播至图像块的梯度均值记为∇L。复合型触发器根据∇L 的正负进行更新。两部分触发器的更新方式一致，下文省去上标 U 和 I。不失一般性，以某一块为例，将 $\boldsymbol{\delta}$ 初始化为全 $\delta_{\min}$或全 $\delta_{\max}$。更新规则如下：

1)若$\nabla L\geqslant 0$ 且 $\boldsymbol{\delta}$ 全为 $\delta_{\min}$，则将 $\boldsymbol{\delta}$ 更新为全 $\delta_{\max}$；

2)若$\nabla L\geqslant 0$ 且 $\boldsymbol{\delta}$ 全为 $\delta_{\max}$，则保持不变；

3)若$\nabla L<0$ 且 $\boldsymbol{\delta}$ 全为 $\delta_{\min}$，则保持不变；

4)若$\nabla L<0$ 且 $\boldsymbol{\delta}$ 全为 $\delta_{\max}$，则将 $\boldsymbol{\delta}$ 更新为全 $\delta_{\min}$。

注意：上述初始化和更新规则均只考虑 $\boldsymbol{\delta}$ 中有定义的位置，即 $m_i=1$。可见，更新过程仅改变余数的取值，块内像素值始终符合同余规律。

3.5 联合视觉-文本特征的损失函数

本文联合视觉-文本特征，通过构建对比损失函数来训练触发器。训练的目标是将异类的文本特征融入复合型触发器之中，进而起到弱化图像内容特征的作用。

具体地，本文引入预训练的 CLIP 模型并冻结其参数。CLIP 模型由 4 亿多对图像文本训练而得，对未见样本具有较好的泛化性，能够在对齐的隐空间中成对地提取视觉和文本特征。由于受害模型的知识未知，本文构建代理模型并进行预训练，将其作为梯度反向传播的桥梁。

将数据集的标签与固定格式的前缀提示词“a photo depicts”拼接后作为 CLIP 文本编码器的输入，得到文本特征集合$\boldsymbol{W}_{\mathrm{txt}}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_K\}$，其中 $\boldsymbol{w}_k$ 为第 k 类标签的文本特征，$1\leqslant k\leqslant K$。将待攻击目标类别的干净图像作为 CLIP 图像编码器的输入，得到图像特征集合$\boldsymbol{V}_{\mathrm{img}}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{N_k}\}$，其中 N_k 表示第 k 类的样本总数。计算目标类别图像的平均特征：

$$\bar{\boldsymbol{v}}_{\mathrm{img}}=\frac{1}{N_k}\sum_{i=1}^{N_k}\boldsymbol{v}_i \quad (5)$$

接着，利用余弦相似距离，从$\boldsymbol{W}_{\mathrm{txt}}$ 中检索出与 $\bar{\boldsymbol{v}}_{\mathrm{img}}$ 相似度最低的异类文本特征$\boldsymbol{w}_{\mathrm{txt}}$：

$$\boldsymbol{w}_{\mathrm{txt}}=\underset{k}{\mathrm{argmin}}\ \mathrm{cs}(\bar{\boldsymbol{v}}_{\mathrm{img}},\boldsymbol{w}_k) \quad (6)$$

其中，$\mathrm{cs}(\cdot,\cdot)$表示余弦相似度的计算。然后，将干净图像 $\boldsymbol{x}$ 和中毒图像 $\boldsymbol{x}+\boldsymbol{\delta}$ 作为预训练代理模型的输入，分别得到视觉特征 $\tilde{\boldsymbol{v}}$ 和 $\tilde{\boldsymbol{v}}'$。

本文构建了基于视觉-文本多模态特征的对比损失函数：

$$L=\|\tilde{\boldsymbol{v}}'-\boldsymbol{w}_{\mathrm{txt}}\|_2^2-\|\tilde{\boldsymbol{v}}-\boldsymbol{v}\|_2^2+[\alpha-\|\tilde{\boldsymbol{v}}'-\tilde{\boldsymbol{v}}\|_2]_+ \quad (7)$$

其中，$\alpha>0$ 表示干净图像和中毒图像之间特征分离距离的裕度值；$[c]_+=\max(0,c)$。式(7)中第一项迫使中毒图像的视觉特征 $\tilde{\boldsymbol{v}}'$与异类文本特征$\boldsymbol{w}_{\mathrm{txt}}$接近；第二、三项分别迫使中毒图像的视觉特征 $\tilde{\boldsymbol{v}}'$与干净图像的视觉特征 $\tilde{\boldsymbol{v}}$(来源于代理模型)和 $\boldsymbol{v}$(来源于 CLIP 的图像编码器)远离。

根据 3.4 节所述的更新规则调整复合型触发器，使得式(7)的损失函数达到极小。更新后的复合型触发器吸收异类的图像和文本特征，有利于弱化中毒图像的内容特征，增强触发器的有效性。

4　实验

本章按照威胁模型的设定，在3个后门攻击领域常用的数据集上向4个经典DNN图像分类模型开展联合视觉-文本特征的复合型触发器净标签后门攻击实验。通过定量和定性实验结果的比较，验证所提方法的触发器在有效性和隐蔽性方面的优势。进一步地，测评所提方法对4种后门防御技术的抵御能力。最后，通过消融实验验证通用型触发器、个性化触发器，以及CLIP模型在后门攻击中发挥的作用和效果，并验证触发器的黑盒迁移能力。

4.1　实验设置

4.1.1　数据集

本文在3个后门攻击领域常用的数据集CIFAR-10[38]，ImageNet[39]和GTSRB[40]上进行实验。CIFAR-10数据集包含飞机、汽车、青蛙等10个类别。每个类包含6000幅32×32的RGB图像，其中5000幅用于训练，1000幅用于测试。为了进行公平的比较，本文按照文献[17,20-23,25]的做法，从ImageNet数据集中随机地选取10个类别，每个类别中的前5000幅图像用于训练，后1500幅图像用于测试。将ImageNet数据集的图像统一缩放为224×224。GTSRB为交通标志数据集，包含43个类别。该数据集各类图像数量不均，最少的包含171幅图像，最多的包含1767幅图像。本文按大约7∶3的比例划分训练集和测试集，并将所有图像统一缩放为48×48。

4.1.2　受害模型

本文将VGG11[41]，ResNet18[42]，DenseNet121[43]和WRN[44]4个经典DNN图像分类模型作为受害模型和代理模型，开展后门攻击实验。若无特别说明，本文默认以DenseNet121作为受害模型和代理模型。对于受害和代理两模型的结构相异的情况，在消融实验中对黑盒触发器的跨模型攻击能力进行验证。

4.1.3　实现细节

本文将各数据集中的首类($k=1$)作为待攻击的目标类别，并将默认的中毒率设置为2%，即中毒图像数量仅占数据集总图像数量的2%。本文采用Sobel算子检测图像的边缘，形成个性化触发器的二进制取样掩码。将个性化触发器的超参数n设置为1，图像的分块大小设置S为16。

本文采用SGD(Stochastic Gradient Descent)[45]作为训练模型的优化器，并将动量设置为0.9。对于CIFAR-10和GTSRB数据集，本文将批大小(Batch Size)设置为128，训练回合数(Epoch)设置为200，SGD的权重衰减设置为5×10^{-4}，学习率初始化为0.1。在第60，120和160个训练回合后，递次地对学习率进行衰减，衰减因子为0.2。对于图像尺寸较大的ImageNet数据集，本文将批大小设置为40，训练回合数设置为60，SGD的权重衰减设置为5×10^{-3}，学习率初始化为0.01。每10个回合对学习率折半衰减。

4.1.4　评价指标

本文采用攻击成功率(Attack Success Rate，ASR)和模型准确率下降(Model Accuracy Decline，MAD)来衡量后门攻击的有效性。前者指成功使得后门模型输出目标标签的中毒图像所占的比例。后者指模型在后门攻击前后，对于干净图像预测准确率的下降量[9]。具体定义如下：

$$ASR=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(f(\boldsymbol{x}_i+\boldsymbol{\delta};\boldsymbol{\theta}')=k) \tag{8}$$

$$MAD=\frac{1}{N}\sum_{i=1}^{N}[\mathbf{1}(f(\boldsymbol{x}_i;\boldsymbol{\theta})=y_i)-\mathbf{1}(f(\boldsymbol{x}_i;\boldsymbol{\theta}')=y_i)] \tag{9}$$

其中，$\mathbf{1}(\cdot)$为指示函数；$\boldsymbol{\theta}$和$\boldsymbol{\theta}'$分别为训练后正常模型和后门模型的参数。ASR在0到$\mathbf{1}$之间，其值越大越好；MAD在-1到1之间，其值越小越好。

本文采用4种常用的图像质量客观评价指标，即峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)[46]、结构相似性度量(Structural Similarity Index Measure，SSIM)[47]、梯度幅度相似性偏差(Gradient Magnitude Similarity Deviation，GMSD)[48]和学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)[49]，从像素、纹理、边缘和特征层面来衡量中毒图像相对于干净图像的视觉质量差异。

4.2　攻击性能的定量比较

本节将所提方法与现有的后门攻击方法(包括BadNets[11]、INK[16]、LCBA[18]、SIG[19]、Refool[20]、Inv[21]、隐蔽图像后门攻击[22]、HTBA[23]、SAA[24]和DA[25])进行定量的性能比较。其中，BadNets、LCBA、SIG、Refool、Inv和隐蔽图像后门攻击的触发器均采用单一固定模式；INK、HTBA、SAA和DA的触发器属于样本特定模式。具体详情如表1所列。此外，BadNets与INK属于脏标签后门攻击。对于这两种攻击，其在净标签设置下的攻击性能以“*”表示。所有方法均按照4.1.3小节所述的设置实现。表2列出了所有方法在CIFAR-10，ImageNet和GTSRB上关于触发器有效性和隐蔽性的定量结果，其中以虚线作为脏标签设置和净标签设置的分割线。本文方法在3个数据集上均取得了最高的ASR，分别为0.9544，0.9995和0.9806，平均高出次优方法2.48个百分点。其中，针对现有方法在GTSRB数据集上类别分布不均衡导致的攻击成功率下降，以及在ImageNet数据集上因图像尺寸较大而削弱隐蔽性的问题，本文方法均展现出了更强的适应性，在攻击有效性和隐蔽性两方面均表现出较优的性能。虽然部分方法(如LCBA、Inv和隐蔽图像后门攻击)取得了0.95以上的ASR，但它们的触发器属于单一固定模式或视觉可见模式，隐蔽性不佳。BadNets与INK仅在脏标签设置下表现较为优异。本文方法在3个数据集上的MAD分别为0.0001，-0.0001和0.0065，均未超过正常模型识别准确率的1%。这意味着后门的植入几乎不影响干净图像的识别准确率。上述结果表明复合型触发器具有较强的有效性。

本文方法在3个数据集上均取得了最佳的PSNR，SSIM，GMSD和LPIPS。其中，PSNR均值为47 dB，分别比Refool，LCBA和Inv高出32 dB，25 dB和33 dB。现有方法中，隐蔽图像后门攻击利用信息隐藏技术合成了较高视觉质量的中毒图像。在上述4项指标上，本文方法平均高出隐蔽图像后门攻击10.61%，0.312%，68.44%和46.38%。上述结果表明复合型触发器不但特定于干净图像，还呈现出信息隐藏的特点，具有较好的隐蔽性。

表 2 不同攻击方法的触发器有效性与隐蔽性的定量对比

Table 2 Quantitative comparison of the trigger effectiveness and stealthiness across different attack methods

数据集	攻击方法	*ASR*↑	*MAD*↓	*PSNR*↑	*SSIM*↑	*GMSD*↓	*LPIPS*↓
CIFAR-10	BadNets[11]	1.0000	0.0104	23.2755	0.9833	0.0459	0.0015
	INK[16]	0.9992	0.0001	42.9515	0.9961	0.0020	0.0002
	BadNets*[11]	0.5751	**0.0001**	23.2755	0.9833	0.0459	0.0015
	INK*[16]	0.5520	**0.0001**	42.9515	0.9961	0.0020	0.0002
	LCBA[18]	0.9121	0.0043	21.6434	0.8762	0.0673	0.0214
	SIG[19]	0.4580	0.0126	23.6781	0.7557	0.1427	0.0495
	Refool[20]	0.8361	0.0080	15.7410	0.6054	0.2366	0.0935
	Inv[21]	0.8932	0.0080	13.3719	0.8460	0.0918	0.5269
	隐蔽图像后门攻击[22]	0.8300	0.0115	43.4324	0.9963	0.0017	0.0001
	HTBA[23]	0.6787	0.0010	21.4134	0.7130	0.0800	0.0418
	SAA[24]	0.9060	0.0013	27.1027	0.9402	0.0172	0.0050
	DA[25]	0.8222	0.0009	21.8243	0.9467	0.0195	0.0476
	本文方法	**0.9544**	**0.0001**	**47.9571**	**0.9992**	**0.0005**	**0.0001**
ImageNet	BadNets[11]	1.0000	0.0020	23.9212	0.9277	0.0806	0.0607
	INK[16]	0.9848	0.0012	40.1223	0.9875	0.0024	0.0050
	BadNets*[11]	0.8020	0.0010	23.9212	0.9277	0.0806	0.0607
	INK*[16]	0.7120	0.0015	40.1223	0.9875	0.0024	0.0050
	LCBA[18]	0.9490	0.0057	23.3435	0.7691	0.1021	0.2594
	SIG[19]	0.5045	**−0.0010**	23.9696	0.8709	0.1066	0.1134
	Refool[20]	0.9599	0.0049	13.7003	0.6327	0.2474	0.4786
	Inv[21]	0.9820	0.0010	17.9664	0.3231	0.0831	0.3968
	隐蔽图像后门攻击[22]	0.9019	0.0040	43.6416	0.9923	0.0018	0.0034
	HTBA[23]	0.6300	0.0001	26.5326	0.7004	0.0981	0.5709
	SAA[24]	0.7851	0.0008	27.2933	0.8391	0.0372	0.1438
	DA[25]	0.8832	0.0005	27.6373	0.9376	0.0163	0.0951
	本文方法	**0.9995**	−0.0001	**46.0899**	**0.9975**	**0.0005**	**0.0005**
GTSRB	BadNets[11]	1.0000	0.0008	24.1426	0.9925	0.0858	0.0283
	INK[16]	0.9781	0.0008	38.0515	0.9965	0.0021	0.0005
	BadNets*[11]	0.4780	0.0083	24.1426	0.9925	0.0858	0.0283
	INK*[16]	0.4037	0.0023	38.0515	0.9965	0.0021	0.0005
	LCBA[18]	0.8972	0.0066	23.6999	0.9514	0.0673	0.0217
	SIG[19]	0.8943	0.0048	21.8946	0.7646	0.1734	0.0563
	Refool[20]	0.7703	**0.0020**	16.7609	0.6162	0.1629	0.2398
	Inv[21]	0.8601	0.0043	13.2614	0.7870	0.1142	0.0824
	隐蔽图像后门攻击[22]	0.9660	0.0033	43.3921	0.9980	0.0008	0.0002
	HTBA[23]	0.6122	0.0058	28.7642	0.9209	0.0203	0.0354
	SAA[24]	0.8155	0.0023	25.5863	0.9111	0.0457	0.0517
	DA[25]	0.7921	0.0022	27.3715	0.9392	0.0193	0.0176
	本文方法	**0.9806**	0.0065	**47.8999**	**0.9992**	**0.0003**	**0.0001**

注：粗体表示最优结果；下划线表示次优结果；符号“↑”(“↓”)表示数值越大(越小)越好。

4.3 中毒图像的视觉质量

图 2 展示了来自 3 个数据集的 6 幅干净图像，以及部分由上述方法合成的中毒图像。通过观察可知，LCBA，SIG，Refool，Inv，HTBA，SAA 这 6 种方法对干净图像的干扰较大，在中毒图像中残留了可见的异类图形、带状条纹、方形色块或周期性噪声等。这容易引起模型开发者的注意，暴露攻击意图。DA、隐蔽图像后门攻击和本文方法对干净图像的修改不大，所合成的中毒图像与干净图像在视觉感知上基本保持了一致。本文所提出的复合型触发器基于像素值的同余规律，排除了对强干扰触发模式的依赖，获得了较好的隐蔽性。

进一步地，本文利用 Elo 评分系统，由人工来评测各种方法所合成中毒图像的主观视觉质量。具体地，将图 2 中所列的 10 种方法视为评测的参与者。其中，恒等变换作为将干净图像合成为干净图像本身的方法。

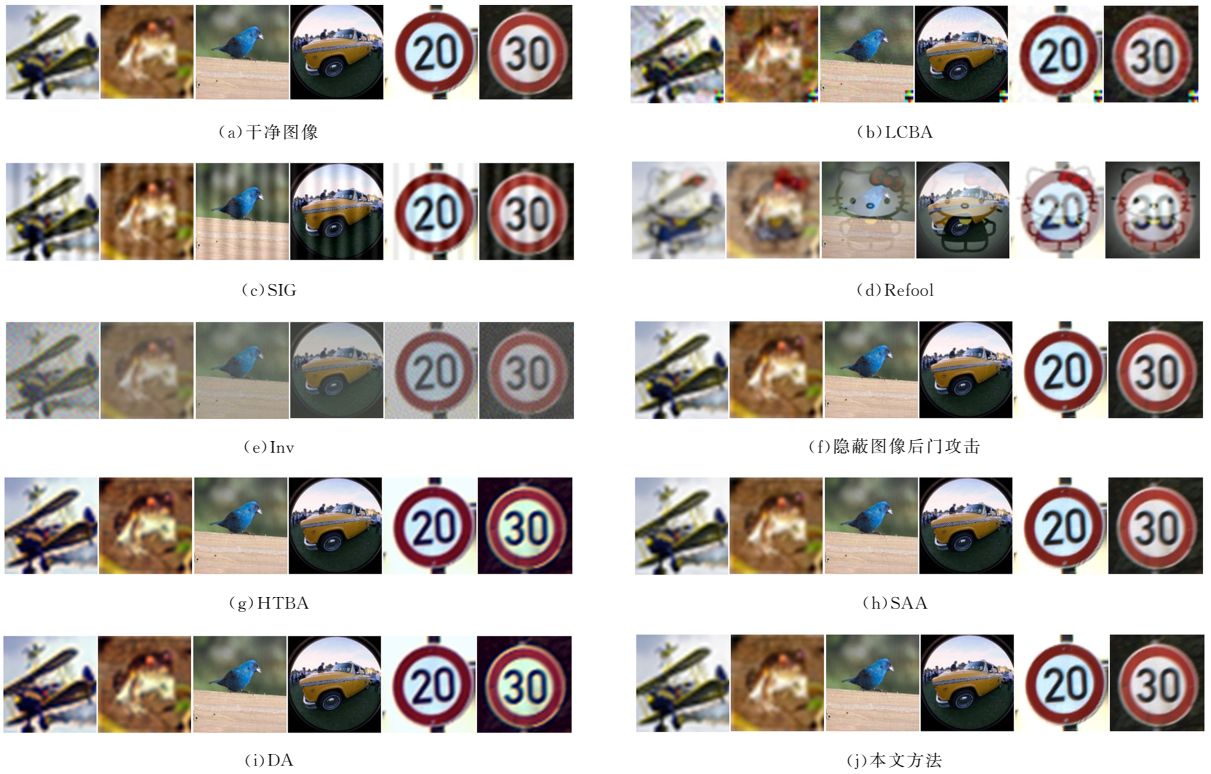
本文邀请了 40 位受试者，每位受试者完成 2 轮 10 种方法中毒图像两两对决的视觉质量对比，共计 90 组。对于每一组评测，将两幅内容相同但由不同方法所合成的中毒图像并排显示在屏幕上。受试者观察两幅中毒图像，进行视觉比对，在 10 秒钟内从“胜”“负”“平”3 个选项中做出选择。

本文按照 Elo 评分系统的标准来执行积分初始化、胜负概率预测和积分更新。初始时，Elo 评分系统为 10 种参评方法设置相同的积分。两两对决前，根据积分预测胜负概率，积分高的方法胜率高。将每一组人工评测的结果转化为概率形式：1.0 表示“胜”，0.0 表示“负”，0.5 表示“平”。结合预测的和实际的胜负概率来更新积分。预测胜率低而实际获胜的方法将获得大量积分。更多的评测细节和示例详见网络[1)]。

图 3 展示了 10 种方法的归一化 Elo 积分排名。本文方法排名第二，仅次于由恒等变换所获得的干净图像。隐蔽图像后门攻击、DA 与本文方法接近，分别排名第三和第四。虽然 LCBA 和 Inv 展现出较高的攻击成功率，但仅排名第七和第九。图 3 所示的排名与表 2 所列的定量结果以及图 2 所示

1) https://github.com/Cyttyc/image-quality

的定性结果一致，进一步验证了本文提出的复合型触发器具有较好的隐蔽性。



(a)干净图像　(b)LCBA　(c)SIG　(d)Refool　(e)Inv　(f)隐蔽图像后门攻击　(g)HTBA　(h)SAA　(i)DA　(j)本文方法

图 2　不同后门攻击方法合成的中毒图像对比

Fig. 2　Comparison of poisoned images synthesized by different backdoor attacks

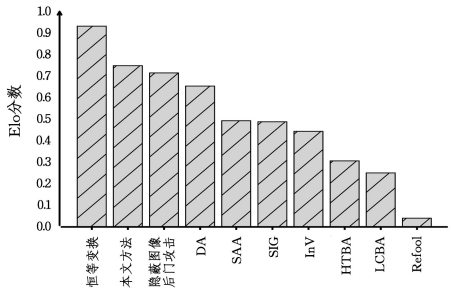


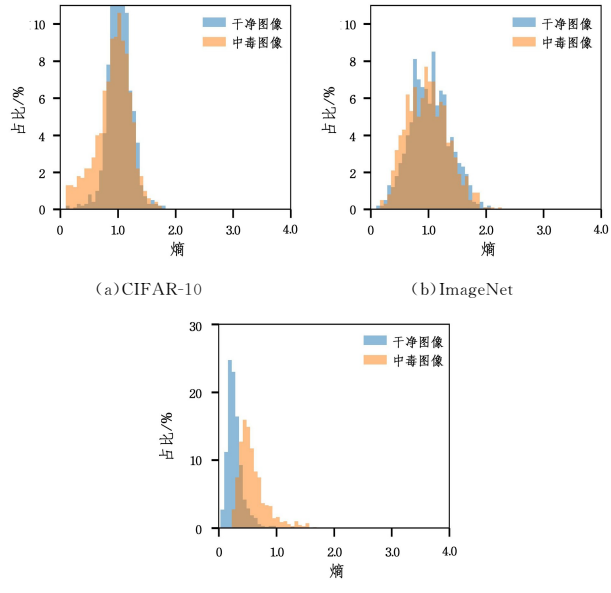
图 3　不同后门攻击方法基于 Elo 评分系统排名

Fig. 3　Ranking results of different backdoor methods based on Elo rating system

4.4　抵御后门防御技术

按照文献[16，22-23，25]的实验设置，本节测评所提方法对 STRIP[28]、SentiNet[29]、神经元清洗[36]和 I-BAU[37]这 4 种后门防御技术的抵御能力。

STRIP 在输入图像中叠加扰动，并监测扰动对后门模型 softmax 层熵的影响，从而识别中毒图像。图 4 展示了在 3 个数据集上基于 STRIP 所获得的熵直方图。对于 CIFAR-10 和 ImageNet 数据集，干净和中毒两类图像的熵直方图高度贴合，表现出较高的一致性。这是由于本文提出的复合型触发器(其中的个性化部分)特定于图像的边缘，起到了防止在 softmax 层中遗留特有痕迹的作用。对于 GTSRB 数据集，两类图像的熵直方图虽略有差异但仍相互交叠且形态相似。这是由于交通标志图像语义明确、背景一致，较难通过叠加扰动来施加影响。上述实验结果表明：所提方法能够较好地应对 STRIP，尤其对于复杂的自然场景图像，防御者无法通过 STRIP 来检测中毒图像。



(a)CIFAR-10　(b)ImageNet　(c)GTSRB

图 4　抵御 STRIP 后门防御的结果

Fig. 4　Results of resisting the STRIP backdoor defense

SentiNet 根据注意力热力图的异常分布来检测中毒图像。图 5 分别展示了 3 个数据集中两幅干净图像和相应的两幅中毒图像的注意力热力图，其中紫红色区域表示后门模型所关注的重点位置。通过比对可知，本文方法所合成的中毒图像并不会显著地改变中毒模型的注意力分布，能够抵御 SentiNet 的检测。

图 5 抵抗 SentiNet 后门防御的结果(电子版为彩图)

Fig. 5 Results of resisting the SentiNet backdoor defense

神经元清洗根据逆向工程所构建的触发器大小来统计绝对中位差异常值。当异常值高于 2 时,可认定受害模型已被植入后门。图 6 展示了在 3 个数据集上训练得到的正常模型和后门模型异常值,图中的红虚线对应报警阈值 2。可以看到,后门模型的异常值均小于 2,与正常模型的表现一致。这说明本文方法能够成功地躲避神经元清洗的排查。

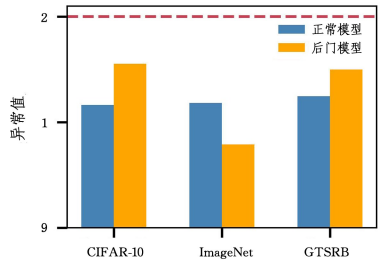
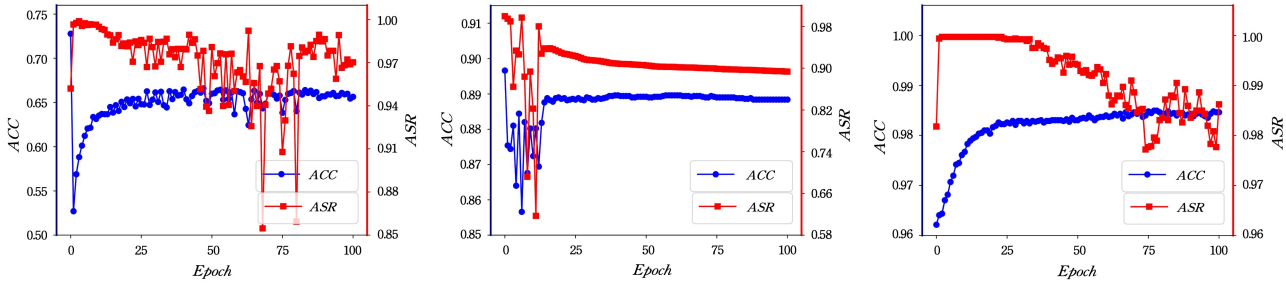


图 6 抵御神经元清洗的防御结果(电子版为彩图)

Fig. 6 Results of resisting the neural cleanse backdoor defense

I-BAU 利用后门遗忘和触发器生成交替训练过程中获得的隐式超梯度来移除后门,实现模型净化。本文遵循文献[37]的设置,采用 SGD 优化器对后门模型净化训练 100 个回合。图 7 展示了模型净化过程中,攻击成功率(ASR)和模型准确率(ACC)的变化趋势。通过观察可知,后门模型的 ACC 基本保持稳定(或略微有所上升)。这是由于模型净化依赖于对干净样本的重学习,有利于保持或提升 ACC。净化 100 个回合后,在 3 个数据集上后门模型的 ASR 仍保持在 0.9 以上。特别地,对于 GTSRB 数据集,净化后的 ASR 仍能达到 0.98,与净化前的 ASR 相比几乎无差别。这是由于本文方法属于净标签后门攻击,且所提出的复合型触发器特定于干净图像,打破了 I-BAU 的防御假设。

综上,上述实验结果验证了本文方法能够较好地抵御触发器检测、后门检测和后门净化等后门防御技术的侦测,展现出较好的隐蔽性。



(a)CIFAR-10 (b)ImageNet (c)GTSRB

图 7 抵抗 I-BAU 后门防御的结果

Fig. 7 Results of resisting the I-BAU backdoor defense

4.5 消融实验

本节围绕黑盒触发器的跨模型攻击能力、中毒率、复合型触发器的设定、触发器分块大小和个性化触发器的超参数 n 开展消融实验,探讨上述 5 个方面对触发器的有效性和隐蔽性的影响。若无特殊说明,本节所有实验均在 CIFAR-10 数据集上进行。

4.5.1 黑盒触发器的跨模型攻击能力

表 3 列出了 VGG11[41],ResNet18[42],DenseNet121[43] 和 WRN[44] 分别作为受害模型和代理模型的各种组合情况下,后门攻击的成功率。由于准确率下降指标 MAD 均保持在 −0.001～0.001 之间,为了避免冗余,未在表 3 中一一列出。

表 3 黑盒触发器的跨模型攻击能力

Table 3 Cross-model attack capability of black-box triggers

代理模型	受害模型			
	VGG11	Resnet18	Densenet121	WRN
VGG11	**0.9992**	0.9688	0.9462	**0.9964**
Resnet18	0.9512	**0.9993**	0.9507	0.9787
Densenet121	0.9880	0.9620	**0.9544**	0.9787
WRN	0.9687	0.9269	0.8721	0.9825

注:粗体表示最优结果。

通过统计可知,表 3 所列的 ASR 平均值为 0.964;跨模型组合情况下,ASR 最高为 0.999,最低为 0.872。相比于同模

型组合，跨模型组合在平均 ASR 指标上仅下降 2.65 个百分点。上述结果表明，经训练后的复合型触发器不但具有较强的有效性，还具有鲁棒的黑盒迁移能力。这一方面得益于复合型触发器中同余规律的显著性，另一方面得益于 CLIP 模型的泛化能力。优秀的跨模型迁移能力为本文方法应用于黑盒攻击场景奠定了基础。

4.5.2　中毒率

将中毒率设为 0.5%，1%，2%，5%和 7%，评测不同中毒率下，后门模型的 ASR 和 MAD，结果如图 8 所示。

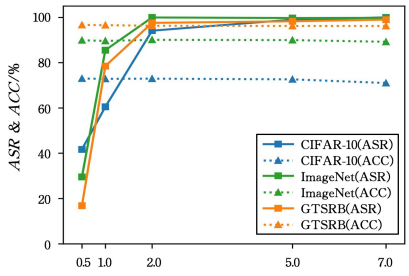


图 8　不同中毒率下复合型触发器性能比较

Fig. 8　Performance comparison of the composite trigger under different poisoning rates

随着中毒率的增加，MAD 略有下降但幅度微小。这表明中毒率小于 7%时，后门的植入几乎不影响干净图像的识别性能。随着中毒率的增加，ASR 先逐步上升后趋于饱和。在中毒率为 1%时，3 个数据集上的 ASR 分别为 0.60，0.85 和 0.71，初步具备了施展后门攻击的可能。当中毒率大于 2%时，3 个数据集上的 ASR 均接近于 1.0。然而，中毒率越高，偏离正常分布的程度越明显，容易暴露攻击意图。综合后门的有效性和隐蔽性，本文将中毒率设为 2%。

4.5.3　复合型触发器的设定

复合型触发器由通用型和个性化两部分叠加而成，并由基于 CLIP 的监督信号驱动训练。考虑“通用型”“个性化”和“CLIP 驱动训练”3 个要素，设置了 6 项消融条件进行测试，结果如表 4 所列。本文采用 ASR 来反映触发器的有效性，采用像素值的最大修改量和神经元清洗异常值来反映触发器的隐蔽性。特别地，ID0 为本文的完整方法，作为消融实验性能比较的基准。

表 4　不同设置下复合型触发器的性能比较

Table 4　Performance comparison of the composite trigger under different settings

编号	通用型触发器	个性化触发器	CLIP 模型	*ASR*	图像修改量	异常值
ID0	√	√	√	**0.9544**	3	1.154
ID1	√	×	√	0.8080	**1**	2.054
ID2	×	√	√	0.4425	2	1.581
ID3	√	△	√	0.8629	3	1.792
ID4	△	√	√	0.7371	3	1.507
ID5	△	△	×	0.8499	3	1.154
ID6	√	√	×	0.8366	3	**0.958**

注：△表示存在但不训练；×表示不存在；√表示存在；“异常值”指神经元清洗基于 MAD 的异常值（该值大于 2 表示从受害模型中检出后门）。

ID1 和 ID2 分别只保留了通用型触发器和个性化触发器。ID1 的 ASR 达到 0.808，显著高于 ID2 的 0.4425。这表明，单一固定的通用型触发器主要在攻击的有效性方面发挥了作用。ID1 的异常值为 2.054，超过了报警阈值 2。这是由于后门模型中的部分神经元反复学习了单一固定的触发器特征，形成了异常的激活状态。相比之下，ID2 的异常值为 1.581，优于 ID1，表明样本特定的个性化触发器主要在攻击的隐蔽性方面发挥作用。复合型触发器结合了两者的优点，同时具备了有效性和隐蔽性。

ID3，ID4 和 ID5 旨在测试复合型触发器（及其中的两部分）的可学习性对后门攻击性能的影响。相比于 ID0 基准，ID3，ID4 和 ID5 在 ASR 指标上分别下降了 9.15 个百分点、21.73 个百分点和 10.45 个百分点。ID4 下降得最多，且低于 ID1 的 ASR。这说明 ID4 所叠加的个性化触发器难以弥补通用型触发器失活（不可学习）带来的损失，从侧面印证了触发器可学习性的重要意义。虽然 ID5 的两部分触发器均不可学习，但是 ID5 仍取得了较高的 ASR 和较低的异常值。这主要归功于两方面：一方面，通用型触发器属于单一固定模式的特性；另一方面，个性化触发器在未训练时，对图像边缘的像素均采用相同的同余规律。因此，即便复合型触发器不可学习，ID5 仍兼顾了有效性和隐蔽性。

ID6 是指虽然通用型和个性化两部分触发器均可学习，但从监督信号中排除了基于 CLIP 模型的多模态特征对比损失项，即式（7）中的第一和第二项。相比于 ID0 基准，ID6 虽然取得了较好的隐蔽性（异常值 0.958 最小），但在 ASR 指标上下降了 11.78 个百分点。这说明，ID0 借助了 CLIP 模型的泛化能力指导触发器学习块间同余规律的组合方式，从而提升有效性。上述消融实验的结果验证了本文提出的可学习复合型触发器既具有较强的有效性，又具有较好的隐蔽性。

4.5.4　触发器分块大小

为明确触发器的分块大小对有效性的影响，本小节在 CIFAR-10 数据集上对不同分块大小的影响进行实验分析，结果如表 5 所列。

表 5　不同分块大小对触发器有效性的影响

Table 5　Impact of different block sizes on the effectiveness of triggers

数据集	分块大小	*ASR*
CIFAR-10	4×4	0.8298
	8×8	0.8743
	16×16	**0.9544**

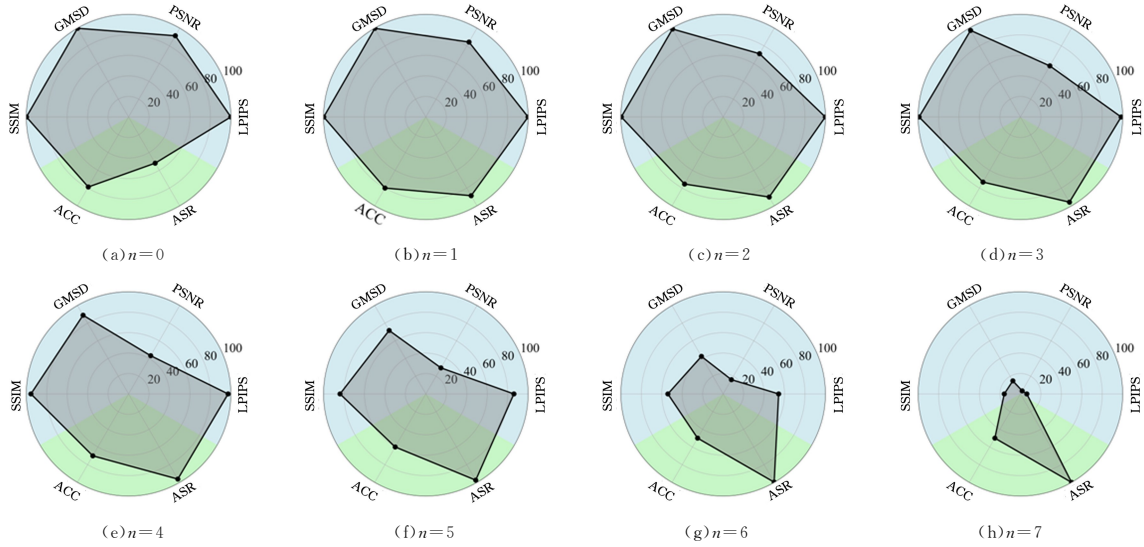
注：粗体表示最优结果。

实验结果表明，当分块大小为 4×4 时，攻击成功率（ASR）仅为 0.829 8，这意味着触发器的有效性较低。当分块大小逐渐增大（如 16×16）时，复合型触发器所具备的同余规律增强，使得受害模型更容易学习触发器特征，从而显著提高攻击成功率（0.954 4）。因此，在实验中，本文采用 16×16 作为分块大小，均匀地将各数据集分为整数块，以便后续受害模型学习，建立触发器与目标标签之间的关联，形成后门。

4.5.5　个性化触发器的超参数

超参数 n，即触发器中同余规律参数，决定了个性化触发器对干净图像边缘像素值的最大修改量，即 $2^{n+1}-2$。图 9 中

的雷达图展示了不同取值 n 对触发器有效性和隐蔽性的影响。特别地，$n=0$ 表示个性化触发器不存在，与表 4 中 ID1 的消融条件一致。雷达图中蓝底色部分涵盖了 PSNR，SSIM，GMSD 和 LPIPS 这 4 项图像质量指标，反映了个性化触发器的隐蔽性；绿底色部分涵盖了 ASR 和模型准确率（ACC）两项指标，反映了个性化触发器的有效性。雷达图中的某个节点代表在该指标下相对于最佳性能值的百分位数。



(a) $n=0$ (b) $n=1$ (c) $n=2$ (d) $n=3$

(e) $n=4$ (f) $n=5$ (g) $n=6$ (h) $n=7$

图 9 个性化触发器在不同超参数 n 下的性能比较（电子版为彩图）

Fig. 9 Performance comparison of individual trigger under different hyperparameters n

通过观察可知，随着 n 增大，包围区域在蓝底色部分逐渐收缩，而在绿底色部分（特别是沿 ASR 轴的方向上）逐渐拓展。这是由于 n 越大，个性化触发器对边缘像素值的修改量越大。这虽然能够提高后门攻击的成功率，但牺牲了中毒图像的视觉质量。当 $n=1$（本文默认设置）时，包围区域的面积最大，说明在中毒图像的视觉质量和后门攻击的成功率之间取得了平衡，很好地兼顾了触发器的有效性和隐蔽性。

结束语 本文探讨了现有后门攻击方法存在着触发器有效性和隐蔽性之间顾此而失彼的矛盾。针对该问题，本文结合单一固定和样本特定两种模式，提出了基于同余规律的复合型触发器净标签后门攻击，兼顾有效性和隐蔽性。复合型触发器由通用型和个性化两部分可学习的触发器叠加而成。通用型触发器属于单一固定模型，其信号形态为 LSB 块状噪声；个性化触发器属于样本特定模型，其信号形态为 LoSB 边缘噪声。复合型触发器诱导受害模型捕捉同余规律，建立触发器和目标标签的关联，形成后门。在此基础上，本文引入预训练的 CLIP 模型，构建联合视觉-文本特征的对比损失，引导复合型触发器吸收异类的文本特征，起到弱化图像内容特征的作用，进一步提高有效性。

本文在 CIFAR-10，ImageNet 和 GTSRB 这 3 个数据集上开展了实验。结果表明，所提方法在 ASR，PSNR，SSIM，GMSD，LPIPS 指标上均超越了现有方法，展现出良好的有效性和隐蔽性。所提方法能够成功抵御神经元清洗、STRIP、SentiNet 和 I-BAU 这 4 种后门防御技术的侦测。消融实验结果表明，所提方法具备跨模型攻击能力，验证了通用型和个性化两部分对触发器的有效性和隐蔽性的贡献。在未来工作中，将结合大语言模型和提示学习，进一步联合多模态特征驱动复合型触发器的训练。同时，在训练阶段引入对抗学习，模拟物理世界中的噪声与失真，提升复合型触发器的鲁棒性和泛化能力。

参 考 文 献

[1] YUAN L, CHEN Y, CUI G, et al. Revisiting Out-of-Distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations[C]// Advances in Neural Information Processing Systems. MIT, 2023: 58478-58507.

[2] CHEN J, TAM D, RAFFEL C, et al. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 191-211.

[3] LENG Y, TAN X, ZHU L, et al. Fastcorrect: Fast Error Correction with Edit Alignment for Automatic Speech Recognition [C]// Advances in Neural Information Processing Systems. MIT, 2021: 21708-21719.

[4] KHEDDAR H, HEMIS M, HIMEUR Y. Automatic Speech Recognition Using Advanced Deep Learning Approaches: A Survey [J]. Information Fusion, 2024, 109: 102422.

[5] YU J, YIN H, XIA X, et al. Self-Supervised Learning for Recommender Systems: A Survey [J]. 2023 IEEE Transactions on Knowledge and Data Engineering, 2023, 36(1): 335-355.

[6] RAJPUT S, MEHTA N, SINGH A, et al. Recommender Systems with Generative Retrieval[C]// Advances in Neural Information Processing Systems. MIT, 2023: 10299-10315.

[7] QIU H, YU B, GONG D, et al. Synface: Face Recognition with Synthetic Data[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 10880-10890.

[8] ZHANG S, GONG Y H, WANG J J. The Development of Deep

Convolution Neural Network and Its Applications on Computer Vision [J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647.

[9] DU W, LIU G S. A Survey of Backdoor Attack in Deep Learning [J]. Journal of Cyber Security, 2022, 7(3): 1-16.

[10] HUANG S X, ZHANG Q X, WANG Y J, et al. Research Progress of Backdoor Attacks in Deep Neural Networks [J]. Computer Science, 2023, 50(9): 52-61.

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

[12] CHEN X Y, LIU C, LI B, et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning [J]. arXiv: 1712.05526, 2017.

[13] NGUYEN A, TRAN A. WaNet-Imperceptible Warping-Based Backdoor Attack [C] // The 9th International Conference on Learning Representations. 2021.

[14] LI S, XUE M, ZHAO B Z H, et al. Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization [J]. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088-2105.

[15] LI Y M, LI Y M, WU B Y, et al. Invisible Backdoor Attack with Sample-Specific Triggers [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 16463-16472.

[16] ZHONG N, QIAN Z, ZHANG X. Imperceptible Backdoor Attack: From Input Space to Feature Representation [C] // Proceedings of the 31th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, 2022: 1736-1742.

[17] ZHANG J, DONGDONG C, HUANG Q, et al. Poison Ink: Robust and Invisible Backdoor Attack [J]. IEEE Transactions on Image Processing, 2022, 31: 5691-5705.

[18] TURNER A, TSIPRAS D, MADRY A. Label-Consistent Backdoor Attacks [J]. arXiv: 1912.02771, 2019.

[19] BARNI M, KALLAS K, TONDI B. A New Backdoor Attack in CNNS by Training Set Corruption without Label Poisoning [C] // 2019 IEEE International Conference on Image Processing. IEEE, 2019: 101-105.

[20] LIU Y, MA X, BAILEY J, et al. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks [C] // The 16th European Conference on Computer Vision. Springer, 2020: 182-199.

[21] NING R, LI J, XIN C S, et al. Invisible Poison: A Blackbox Clean Label Backdoor Attack to Deep Neural Networks [C] // Proceedings of 2021 IEEE Conference on Computer Communications. IEEE, 2021: 1-10.

[22] ZHU S W, LUO G, WEI P, et al. Image-Imperceptible Backdoor Attacks [J]. Journal of Image and Graphics, 2023, 28(3): 864-877.

[23] SAHA A, SUBRAMANYA A, PIRS-IAVASH H. Hidden Trigger Backdoor Attacks [C] // Proceedings of the 34th AAAI Conference on Artificial Intelligence. AAAI, 2020: 11957-11965.

[24] SOURI H, FOWL L, CHELLAPPA R, et al. Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch [C] // Advances in Neural Information Processing Systems. MIT, 2022: 19165-19178.

[25] XU C, LIU W, ZHENG Y, et al. An Imperceptible Data Augmentation Based Blackbox Clean-Label Backdoor Attack on Deep Neural Networks [J]. IEEE Transactions on Circuits and Systems, 2023, 70(12): 2011-5024.

[26] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision [C] // Proceedings of the 38th International Conference on Machine Learning. 2021: 8748-8763.

[27] TRAN B, LI J, MADRY A. Spectral Signatures in Backdoor Attacks [C] // Advances in Neural Information Processing Systems. MIT, 2018: 8011-8021.

[28] 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. IEEE, 2019: 113-125.

[29] CHOU E, TRAMER F, PELLEGRINO G. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems [C] // 2020 IEEE Security and Privacy Workshops. IEEE, 2020: 48-54.

[30] SELVARAJUR R, COGSWELL M, DAS A, et al. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2017: 618-626.

[31] DOAN B G, ABBASNEJAD E, RANASINGHE D C. Februus: Input Purifycation Defense against Trojan Attacks on Deep Neural Network Systems [C] // Annual Computer Security Applications Conference. ACM, 2020: 897-912.

[32] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering [C] // 2019 Proceedings of the Workshop on Artificial Intelligence. AAAI, 2019.

[33] LI Y, LYU X, KOREN N, et al. Anti-backdoor Learning: Training Clean Models on Poisoned Data [C] // Advances in Neural Information Processing Systems. MIT, 2021: 14900-14912.

[34] ZHENG R, TANG R, LI J, et al. Pre-activation Distributions Expose Backdoor Neurons [C] // Advances in Neural Information Processing Systems. MIT, 2022: 18667-18680.

[35] LIU K, DOLAN-GAVITT B, GARG S. Fine-Pruning: Defending against Backdooring Attacks on Deep Neural Networks [C] // Research in Attacks, Intrusions, and Defenses. Springer, 2018: 273-294.

[36] 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. IEEE, 2019: 707-723.

[37] ZENG Y, CHEN S, PARK W, et al. Adversarial Unlearning of Backdoors via Implicit Hypergradient [C] // The 10th International Conference on Learning Representations. 2022.

[38] KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features from Tiny Images. [EB/OL]. [2024-12-13]. https://www.cs.Utor onto.ca/~kriz/learning-features-2009-TR.pdf.

[39] DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.

[40] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition[C]// The 2011 International Joint Conference on Neural Networks. IEEE, 2011: 1453-1460.

[41] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. arXiv: 1409.1556, 2014.

[42] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.

[43] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 4700-4708.

[44] ZAGORUYKO S, KOMODAKIS N. Wide Residual Networks [J]. arXiv: 1605.07146, 2016.

[45] LOSHCHILOV I, HUTTER F. SGDR: Stochastic Gradient Descent with Warm Restarts [C]// The 5th International Conference on Learning Representations. 2017.

[46] HUYNH-THU Q, GHANBARI M. Scope of Validity of PSNR in Image/Video Quality Assessment [J]. Electronics Letters, 2008, 44(13): 800-801.

[47] WANG Z, BOVIK A C, SHEIKH H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[48] XUE W, ZHANG L, MOU X, et al. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index [J]. IEEE Transactions on Image Processing, 2013, 23(2): 684-695.

[49] ZHANG R, ISOLA P, EFROS A A, et al. The Unreasonable Effectiveness of Deep Features as A Perceptual Metric[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 586-595.

HUANG Rong, born in 1985, Ph.D, associate professor, is a member of CCF (No. V6245M). His main research interests include deep neural network model security, multimedia information and computer vision.

(责任编辑:何杨)