密 级 公开

桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

（全日制专业学位硕士）

题目 基于自然触发器的后门攻击方法研究

（英文） Research on Backdoor Attack Method on Natural Trigger

研 究 生 学 号: 19032303064

研 究 生 姓 名: 周 礼

指导教师姓名、职称: 赵 峰 研究员

申 请 学 位 类 别: 工 程 硕 士

领 域: 计 算 机 技 术

论 文 答 辩 日 期: 2022 年 06 月 02 日

独创性（或创新性）声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得桂林电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：周弘　　日期：2022年6月3日

关于论文使用授权的说明

本人完全了解桂林电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属桂林电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为桂林电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本人签名：周弘　　日期：2022年6月3日

导师签名：赵峰　　日期：2022年6月3日

摘要

近年来，深度学习基于在处理大量数据方面的巨大潜力，已经在多个领域取得了重大进展，但同时后门攻击等安全问题也严重威胁深度神经网络模型。后门攻击指攻击者意图在深度神经网络（Deep Neural Networks, DNN）中注入隐藏的后门，使被攻击模型在良性样本上表现良好，并且可通过由攻击者设定好的后门触发器影响模型预测。其研究可以在深度学习安全方面起到重要的作用。

目前，最流行有效的后门触发器往往是在固定位置使用同一个触发器处理不同的干净数据，或嵌入触发器时不考虑宿主数据的内容导致与宿主样本内容相关性差，此外，中毒样本可能是通过简单地将触发器与良性宿主样本直接叠加而生成。因此，由以上触发器生成的后门样本不可避免地存在异常分布，不能自然地将后门嵌入模型，容易引起模型开发人员/用户的怀疑，可能在模型训练阶段之前被过滤掉，或者在模型推理之前被拒绝。另一方面，研究人员努力提高 DNN 模型的鲁棒性，并提出了各种后门对策来去除或抑制 DNN 模型的后门行为。已有研究表明，大多数现有的后门攻击方法都可以通过一些当前流行的防御措施成功地缓解，如微调、精细微调剪枝和基于梯度的类激活图防御措施。因此针对上述问题，本文主要开展了以下两个工作：

（1）提出了一种用于图像分类任务的后门攻击：基于雨滴触发器的后门攻击方法（RDBA）。首先通过随机噪声与 α 值保证雨滴触发器的随机分布以及控制触发器密度，再通过构建好的对角矩阵与旋转矩阵进行仿射变换得到一个对角核，最后对该核进行高斯模糊且使用该核对初步生成的噪声图进行滤波操作，使得触发器完成由从最初的随机噪声图到有宽度、长度、运动模糊的雨滴图的转变。然后，将雨滴触发器与一小部分干净的训练样本合并，生成触发器与宿主样本融合度高的有毒数据。最后，通过多分类网络模型训练得到后门攻击模型，并基于 ImageNet 和 GTSRB 数据集进行实验验证，证明了该方法在使用当前流行的防御机制缓解后门感染模型的后门行为时的鲁棒性、有效性、隐秘性等。

（2）提出了一种基于图像隐写触发器的后门攻击方法（ISBA）。首先，从干净数据集中取出一小部分样本，对每一个样本利用二维离散傅里叶变换进行空间域到频域的转换操作，接着将由攻击者制作的带有目标类别信息的触发器图像与以上频域样本结合，再通过逆二维离散傅里叶变换将已被嵌入触发器的频域样本转换回空间域得到后门攻击样本。然后，将后门样本与干净样本混合共同送入多分类神经网络训练，以得到后门模型。最后，在 ImageNet 和 GTSRB 数据集上用实验证明了该方法在面对当前流行的防御机制攻击模型时的有效性、鲁棒性等。

关键词：后门攻击；深度学习；自然行为；雨滴；图像隐写；隐秘性；攻击有效性

Abstract

In recent years, due to the great potential of deep learning in processing large amounts of data, it has made great progress in many fields, but at the same time, security problems such as backdoor attacks also seriously threaten the deep neural network model. Backdoor attack means that the attacker intends to inject hidden backdoors into DNN (Deep Neural Networks, DNN) to make the attacked model perform well on benign samples and influence model predictions through backdoor triggers set by the attacker. The research of backdoor attack can play an important role in deep learning security.

Currently, the most popular and effective backdoor triggers tend to use the same trigger in a fixed location to handle different clean data, or the contents of host data are not considered when the trigger is embedded, resulting in poor correlation with the contents of the host sample. In addition, the poisoned sample may be generated by simply superposing the trigger directly with the benign host sample. Therefore, the backdoor samples generated by the above triggers inevitably have abnormal distribution and cannot be naturally embedded into the model, easily arousing the suspicion of the model developer/user and may be filtered out before the model training stage or rejected before the model reasoning. On the other hand, researchers have made great efforts to improve the robustness of DNN models and proposed various backdoor strategies to remove or suppress the backdoor behavior of DNN models. Studies have shown that most of the existing backdoor approaches can be successfully mitigated with some of the currently popular defenses such as fine-tuning, fine-tuning pruning, and Grad-CAM based defenses. Therefore, in view of the above problems, this paper mainly does the following two works:

(1) A backdoor attack for image classification task is proposed: Backdoor attack method based on raindrop trigger (RDBA). Firstly, the random noise and α values are used to ensure the uniform random distribution of raindrop trigger and control the density of trigger. Then, a diagonal kernel is obtained by affine transformation of the constructed diagonal matrix and rotation matrix. Finally, Gaussian blur is performed on the kernel and the noise image generated by the check is used for filtering operation. The trigger is transformed from a random noise pattern to a raindrop pattern with width, length and motion blur. Then merged the raindrop trigger with small, clean training samples to produce natural-looking poisoning data. Finally, the backdoor model is obtained by training the multi-classification network model, and the experimental verification based on ImageNet and GTSRB data sets proves

the robustness, effectiveness and stealthiness of the proposed method in alleviating the backdoor behavior of the backdoor infection model by using the current popular defense mechanism.

(2) A backdoor attack method based on image steganography trigger (ISBA) is proposed. Firstly of all, a small number of samples were taken from the clean data set, and each sample was converted from the spatial domain to the frequency domain using the 2D discrete Fourier transform. Next, the trigger image with target category information made by the attacker was combined with the above frequency domain samples. After that, the backdoor attack sample is obtained by converting the frequency domain sample which has been embedded into the trigger back to space domain by inverse two-dimensional discrete Fourier transform. Then, the backdoor samples and clean samples are mixed and sent to multi-classification neural network training to obtain the backdoor model. Finally, the experimental verification based on ImageNet and GTSRB data sets proves the effectiveness and robustness of the proposed method against the current popular defense mechanism attack models.

Keywords: backdoor attack; deep learning; natural behavior; raindrops; image steganography; stealthiness; attack effectiveness

目录

第一章 绪论

§1.1 课题研究背景与意义

如今深度学习技术已被证明在处理海量高维数据方面非常有效，目前已被广泛应用于各种领域，如智能驾驶[1][2][3]、计算机视觉[4][5]、自然语言处理[6][7][8]等。随着深度神经网络向更深更宽的架构演进，导致需要学习的参数变多，同时消耗更多的计算资源才能完成各种预期任务。谷歌、亚马逊等各大公司都推出了自己的机器学习服务（Machine Learning as a Services, MLaaS）平台，方便用户外包模型培训项目。因此，深度神经网络（Deep Neural Networks, DNN）能否被训练到足以完成预期任务与否基本取决于网络开发者所拥有的算力和训练数据。有时用户存在数据集搜集困难或者是控制训练成本问题，用户可以选择采用第三方数据集，还可以基于第三方平台训练DNN，而不是在本地训练 DNN，甚至可以直接使用第三方预先训练的模型。便利背后的成本是对训练阶段的控制权丧失，这可能会进一步扩大训练的安全风险。同时，深度神经网络的安全漏洞可能出现在供应链的任何阶段，包括但不限于未受保护的开放渠道、不可靠的数据源和不可靠的训练过程。研究表明，深度神经网络容易受到不同阶段的攻击，包括推理阶段攻击[9][10][11]和训练阶段攻击。推理阶段攻击以对抗性攻击最为著名[12][13][14][15]，其通常旨在误导深度神经网络，在推理过程中对测试数据产生高置信度的误差预测结果。它们通常是通过在查询目标模型之前向测试数据添加微小的特定扰动来实现的。

训练阶段攻击通常是指后门攻击[16]，其目的是通过毒害一些正常的训练数据来操纵由攻击者生成的后门攻击样本的模型预测。具体来说，攻击者将一些模式（称为触发器）嵌入到干净的图像示例中。这些修改后的数据也被称为中毒样本或后门样本，它们被分配预先定义的目标标签。通过同时引入后门样本和正常数据进行模型训练，从而使训练后的模型被嵌入后门。在推理阶段，后门模型对良性输入正常执行预测，但对包含触发器实例则预测由攻击者选择的目标标签。这种类型的攻击一定程度上是隐秘的，因为后门感染模型在干净输入上有着良好的性能，这类性能与它对应的干净模型是无法区分的，而它的后门行为只能由攻击者指定的（未知的）输入激活。故后门攻击旨在将隐藏的后门嵌入到深度神经网络中，以便后门感染模型在良性样本上表现良好，而一旦隐藏的后门被攻击者指定的触发器激活，它们的预测将被恶意改变。这构成了新的和现实的威胁，当训练过程没有被完全控制时，例如在第三方数据集上进行训练或采用第三方模型，则可能会发生这种威胁。总而言之，这类潜在的恶意行为可能会给一些安全关键领域造成严重的后果，如自动驾驶[17]、人脸识别[18][19][20]、

语音识别[21][22][23]，这也会给 DNN 的部署和发展带来严重的障碍。

在文献中，最流行有效的后门触发器往往是在固定位置使用同一个触发器处理不同的干净数据，或者嵌入触发器时不考虑宿主数据的内容导致与宿主样本内容相关性差。此外，中毒样本可能是通过简单地将触发器与良性宿主样本直接叠加而生成的。例如有的触发器只是图像右下角的一个简单黑白像素块。这些后门样本数据将不可避免地有异常的分布，看起来不自然，容易引起模型开发人员/用户的怀疑。它们可以很容易地在模型训练阶段之前被过滤掉，或者在模型推理阶段之前被拒绝。另一方面，研究人员为提高 DNN 模型的鲁棒性做出了巨大的努力，并提出了各种后门防御对策来去除或抑制 DNN 模型的后门行为。已有研究表明，大多数现有的后门方法其后门行为都可以通过一些当前流行的防御措施成功地缓解，如微调防御[24]、精细剪枝防御[25]和基于梯度的类激活图防御措施[26][27]。

§1.2 后门攻击国内外研究现状

后门攻击指的是一种可以通过在训练或微调过程中毒害一小部分干净的数据集，对遇到特定输入时的模型预测结果进行恶意操作的技术。后门攻击已经对模型供应链造成了严重的威胁，并引起了业界和研究界的广泛关注。具体来说，攻击者将精心制作的触发器嵌入到一些良性的训练数据中，以创建有毒的后门样本。在推理阶段中，针对这些有毒数据训练的神经网络将在输入包含触发器信息时激活异常行为，但在遇到良性输入时却表现正常。

§1.2.1 不同标签设置下的后门攻击

现有的后门攻击根据中毒样本的标签是否被改变，可以分为有毒标签攻击和干净标签攻击。基于有毒标签的攻击是在训练前将中毒数据的标签替换为由攻击者预定义的目标标签。因此，当后门模型检测到触发器时，它的预测将是目标标签。BadNets[16]是在机器学习训练阶段揭示后门威胁的最流行的作品之一。作者使用图像右下角的一个简单的二进制像素块作为后门触发器，通过将触发器嵌入一个良性实例并将其标签更改为目标标签来创建中毒样本。然而，由于触发器相对于其宿主样本而言是异常值，这类攻击很容易被人工检测或后门检测机制检测到。作为改进，Chen 等人在文献[28]中将触发器与良性图像混合生成有毒图像，如使用 Hello kitty 图像作为触发器图像与干净的样本重叠，触发器具有一定的透明度。然而，上面讨论的触发器均是固定位置上的固定模式，正如第 1.2.3 节所描述，大多数现有的有毒标签后门攻击行为都很容易被当前流行的防御措施所缓解。

在基于干净标签的后门攻击中，中毒的训练数据样本仍然保留它们的原本的真实

标签即源标签，它们看起来像输入空间中的源干净样本或像素级的源干净样本。一个典型的工作是由 Shafahi 等人[29]提出的，作者通过在训练集的目标类的干净样本中添加不明显的扰动来制作有毒数据。被污染的数据看起来像干净的数据样本（即来自测试集的某个非目标类的干净数据），然而，在潜在的特征空间中，它们却更接近标签为目标类别的样本。在推理过程中，该类输入将被后门感染模型误分类为目标类。Zhu 等人[30]指出，Shafahi 等人[29]提出的工作不适用于黑盒设置，因为一般受害者网络参数是不可访问的。故他们提出了改进版本，利用凸多面体攻击来制作中毒样本。由于中毒样本的内容与其标签一致，这些数据即使经过人检也会被认为是良性样本。因此，基于干净标签的攻击比有毒标签攻击更隐秘。但是，可能是因为触发器是一组特定的测试数据，而不是一个通用的模式，所以基于干净标签的后门攻击其成功率相对较低，例如 Shafahi 等人[29] 提出的方法，对于一个 10 类分类设置，攻击成功率为 60%。因此，在本工作中，我们只针对有毒标签攻击，提出了两种基于自然触发器的后门方法。

§1.2.2 不同场景下后门攻击

如 1.1 节中所说，在实际生活的运用中，深度学习网络在物联网等多个领域下的应用场景是十分复杂的，同时，深度学习理论处在持续发展趋势下，这在不同的场景中创造了神经网络的不同特征。深度神经网络能被训练到足以完成预期任务与否基本是取决于网络开发者能否拥有获取足够算力和训练数据。有时用户存在数据集搜集困难或者是控制训练成本问题，这时用户往往可以选择采用第三方数据集，而不是自己收集训练数据，因为互联网上有许多免费可用的数据集; 用户还可以基于第三方平台训练 DNN，而不是在本地训练 DNN；用户甚至可以直接使用第三方预先训练的模型。便利背后的成本是对训练阶段的控制权丧失，这可能会进一步扩大训练的安全风险。总之，深度神经网络的安全漏洞可能出现在供应链的任何阶段，包括但不限于未受保护的开放渠道、不可靠的数据源和不可靠的训练过程。因此，根据不同的目标场景，神经网络后门的攻击可以应用于几种不同的攻击策略。

（1）外包训练。由于深度神经网络特的学习训练需要对网络中数以百万计的权重值进行调整，这需要引入大量的优质数据参与训练过程才能得到一个性能良好的模型，因此对网络模型进行训练时需要大量消耗计算资源。对大部分的开发者来说，为了节省训练开销，将网络模型的训练交由其他提供外包的服务商才能节省自身资源。在此类场景下，若攻击者作为被开发者选中的服务商，其可以拥有对模型训练过程的完全控制权，此时后门可通过不受限制的方式被攻击者嵌入网络模型。基于 BadNets[16]的后门攻击是针对该类场景提出的。

（2）迁移学习。迁移学习的初始目的是使为某一任务专门化的预训练模型可以不用从头开始训练就可以重新应用于一个与其初始任务不同但相关的新任务上，而无需

耗费大量时间和资源重新训练网络模型，从而达到节约计算成本的目的[31]。如一网络模型源任务为识别猫的眼睛，新的不同但相关的任务为识别狗的眼睛，此时可以用到迁移学习节省计算资源。如若后门攻击者打算在迁移学习场景下对受害者模型发起攻击，攻击者可以选择直接控制或间接影响预训练模型的再训练过程向进行迁移学习的预训练模型植入后门，如 Trojaning [32]和 Poison[29]。在第一种后门攻击方法中，攻击者对于模型的再训练过程拥有完全控制权，同时攻击者对与训练模型的参数进行访问控制，可以自由对训练数据集访问和处理，以此在对模型再训练过程中嵌入后门。相比于第一个后门攻击方法，Poison 方法中后门攻击者可以通过上传包含中毒样本的训练数据间接影响模型的再训练过程，但整体对模型训练掌控的权限较上一个攻击方法而言较小，无法通过直接对模型参数进行随意的访问和修改来控制模型再训练过程。

（3）深度强化学习。深度强化学习具有决策能力，使强化学习与深度学习相结合。强化学习简单来说是智能体在进行某任务时先于环境交互产生新状态，同时环境给出奖励，循环往复多次之后智能体得以学会完成任务所需要的动作策略。深度强化学习则利用深度神经网络强大的拟合表征的能力与环境交互去拟合动作策略，根据环境对动作的奖励产生新的结果，最终获得最优策略[33]。对于深度强化学习场景，通常使用马尔可夫决策过程，其序贯性和奖励机制让后门获得新特点。Yang 等人[34] 在文献中提出了关于深度强化学习的后门攻击，该方法在应用于深度强化学习时作出了以下假设：首先，后门攻击者无法改变策略和价值网络的架构。其次攻击者无法更改用于训练智能体的强化学习算法。即后门攻击者只能改变智能体和环境之间通信的状态、动作和奖励。而 Kiourti 等人[35]在文献中使用数据中毒的方式以及对操作对应奖励实现高隐秘性的后门攻击方法。

（4）联邦学习。联邦学习可被称作一种机器学习的框架，或可称其为一种分布式机器学习技术，其应用场景为，如因数据隐私问题，医院不愿意共享数据，但训练模型需要很多优质数据才能使得训练出的模型有较高的预测性能，由于诸如医院之类的用户与用户之间信息不共享，为了训练一个性能良好的模型，此时联邦学习派上用场，在保证用户数据隐私前提下，中心服务器可以从不同的用户提取其根据自身数据更新的模型信息如梯度信息，然后中心服务器将这些用户提供的梯度信息进行安全聚合以训练出一个全局的神经网络模型供不同的用户使用。联邦学习一定程度上保证以上参与者的数据建立在安全合规的保密模式下训练[36][37][38][39]。但是虽然联邦学习采用的安全聚合的方式一定程度保证了用户隐私，但却会导致异常检测变得困难。例如，Bagdasaryan 等人[40]利用了在联邦学习中，用户可以先使用自己的数据训练模型梯度信息，此时，如若用户即为后门攻击者，则其可以训练一个后门感染模型再交由中心服务器。同时由于该后门感染模型在安全聚合方式下无法对其进行异常检测，这导致攻击者可以隐秘地将后门嵌入模型并提交给中心服务器。而 Sun 等人[41]在 Mcmahan 等人[36]的工作基础上，允许非恶意用户从目标任务中正确标记样本标签使得后门攻

击的鲁棒性增强。

§1.2.3 后门防御

后门防御的目的是在模型部署之前或之后检测或缓解后门攻击行为。在目前的文献中，已经提出了各种防御技术。在这里，本文列出了在部署 DNN 模型之前的三种主流后门防御技术，这些技术在实践中是最常使用的。

（1）微调防御（Fine-tuning）：在现实生活中，有些深度神经网络任务的数据集是难以获取的或者可以获取到的数据集很少，而当标记的训练数据不足时，要想获得性能良好的 DNN 网络模型，微调是一种实用和轻量级的选择。研究人员发现，深度学习模型在训练一系列新任务时，极大程度上会出现对之前的学习任务产生灾难性的遗忘。使用微调来防御后门攻击的基本原理是，学习新任务通常会导致模型权重的巨大变化，这将破坏之前网络模型所学习到的触发器特征表示[24]。微调防御利用这种灾难性的遗忘现象来驱动后门感染模型忘记自身被嵌入的后门。也就是说，如果防御者使用一些新的干净的训练数据在后门感染模型上或该网络结构的某几层来训练模型，那么结果是后门感染模型可能会逐渐缓解后门行为并最终遗忘后门行为，因为它在微调过程中没有遇到来自新的训练数据集内部的任何触发器模式。然而，在现实中，当提到后门防御时，单独的微调并不总是有效的。这是因为，与后门相关的神经元和与原始任务相关的神经元往往可能是分离的，并不交叉，它们的权值对原始（或新的）任务的贡献很小。换言之，由良性样本激活的神经元和那些由包含触发器的后门样本所激活的神经元往往重叠率非常低。在微调过程中，由于缺乏足够的驱动力，与后门相关神经元的权值保持不变，后门行为仍然存在。

（2）精细剪枝防御（Fine-pruning）：对于微调防御而言，Gu 等人[16]认为，由良性样本激活的神经元和那些由包含触发器的样本激活的神经元往往重叠率低或并不重叠。换句话说，有些神经元只能被触发器激活，当输入是良性样本时，这类神经元会始终保持休眠状态。鉴于此，想要移除这些对触发器敏感的神经元，即后门神经元，正如精细剪枝防御假说表示，可以在不影响正常数据下网络模型在干净输入数据集上的分类性能的情况下使后门神经元失效。然而，Liu 等人[25]进一步发现，被良性输入激活的神经元子集和被恶意输入激活的神经元子集是可能存在重叠的。后门防御也可以通过抑制被良性输入激活的神经元来缓解后门行为。但是在这种情况下，单独修剪神经元将不可避免地导致良性输入的性能损失，因为想要通过剪枝达到防御目的，势必修剪的神经元比例越高，其将后门神经元完全移除的可能性越大。考虑到这些缺点，Liu 等人[25]中提出了精细剪枝防御，它将神经元剪枝的优点与微调防御相结合，在移除后门神经元的同时，通过对后门感染模型进行微调来恢复模型在干净输入数据集上的分类性能。

（3）基于梯度的类激活图防御（Grad Class Activation Map, Grad-CAM）：Grad-CAM[26][27]是一种模型决策的可视化解释和对象检测的常用且有用的技术。它通过识别对模型预测结果贡献最大的样本激活区域来生成关于 DNN 决策结果的可视化解释。基于梯度的类激活图防御方法则是主要利用以上技术来识别恶意的显著区域，即触发器所在区域，并由此过滤掉潜在的异常输入或行为。例如 Chou 等人提出的 SentiNet[42]，利用 Grad-CAM 和边界分析来定位每个样本被划分为某一类时的激活区域，即跨不同实例的通用区域。当被定位的跨不同实例的通用区域相似性高时则可能被开发者认定为可疑触发器。然后，该方法通过将显著的激活区域与普通区域分离后将其覆盖于其他干净样本上输入后门感染模型，根据预测结果判定其是否为后门触发器。同时，Huang 等人提出的 NeuronInspect[43]也遵循了这一思路来检测中毒样本。综上所述，基于梯度的类激活图防御的有效性主要依赖于对于触发器在样本中的激活区域的定位来检测后门样本。

§1.3 论文的主要工作及结构安排

本文首先对主流后门攻击相关方法进行了回顾与分析，分析总结了目前这些后门方法存在的不足之处，这些后门攻击任务方法往往是在固定位置使用同一个触发器处理不同的干净数据，或者嵌入触发器时不考虑宿主数据的内容导致与宿主样本内容相关性差。由这类后门方法得到的后门感染模型在实际部署中很容易被开发者拒绝。此外，后门是通过简单地将触发器写入良性宿主样本而生成的。在实际的应用中，这些触发器同样并不能抵抗目前一些主流的后门防御方法对于后门行为的缓解。为此，本文针对以上问题，提出了两种基于自然后门触发器的后门攻击方法，这两类方法可以将后门自然地嵌入受害者模型中：（1）基于雨滴触发器的后门攻击方法研究。（2）基于图像隐写触发器的后门攻击方法研究。以上两类方法在准确性、鲁棒性、攻击有效性等标准上都表现良好，且能够很好平衡后门触发器的隐秘性，从而自然地将触发器嵌入受害者模型中，能够更好地在实际应用中部署。本文的各章节的结构安排如下：

第一章，首先阐述基于图像任务的后门攻击背景和研究意义。之后从两个维度：基于干净标签或基于中毒标签、不同应用场景，概述了传统流行的后门攻击算法。最后，概述了现有部署 DNN 模型之前的主流后门防御技术。

第二章，本章节首先说明了后门攻击的模型与攻击目标，以及后门攻击的常用专业术语定义。然后介绍了本文中使用到的两个卷积神经网络：VGG16 和 ResNet18。最后，在介绍完基础的分类神经网络后，讲解了本文后门攻击任务中用到的数据集。

第三章，提出了基于雨滴现象的后门攻击模型（RDBA）。首先通过对随机噪声进行不同种滤波等操作生成后门触发器雨滴。再通过使用交叉熵损失的标准模型训练过程完成从后门触发器到目标标签的映射学习。最后，在两种不同的网络上分别用不同

的数据集进行多组实验，验证了使用 RDBA 生成不同后门感染模型时在准确性、攻击有效性、隐秘性和鲁棒性四个方面的性能。

第四章，提出了基于图像隐写的后门攻击模型。首先阐述了现有后门攻击方法的不足之处，然后利用基于二维离散傅里叶变换生成后门隐写触发器，并训练后门感染模型。最后在多个数据集与网络上通过多组实验证明图像隐写后门模型的有效性。

第五章，首先对本文所提出后门攻击方法的创新点和解决的问题做出总结。之后，说明对该工作将来可能做出的改进。

第二章 相关背景知识概述

本章节主要是对后门攻击相关知识的阐述，首先说明了本文使用到的后门攻击的专业术语。然后介绍了后门攻击的威胁模型与攻击目标以及后门攻击过程中的实体。接下来介绍本文中使用到的两个卷积神经网络：VGG16 和 ResNet18。在介绍完以上两类基础的图像分类神经网络相关知识后，讲解了在本文的后门攻击任务中使用到的两类数据集。

§2.1 后门攻击模型与目标介绍

§2.1.1 后门攻击专业术语定义

本节将简要描述和解释后门攻击中使用的专业术语。在本文的剩余部分，我们将遵循相同的后门术语定义。

• 后门触发器（触发器）：由攻击者添加到干净训练数据中的某类特征，通过触发器激活后门感染模型中的隐藏后门并映射目标标签。

• 良性模型：是指完全由干净训练数据训练得来的网络模型。

• 后门感染模型：是指通过包含触发器的训练数据得来的具有隐藏后门的模型。

• 源标签：指示中毒或受攻击样本的真实标签。

• 目标标签：目标标签由攻击者指定，表示攻击者打算使所有包含触发器的样本被后门感染模型所返回的标签。

• 后门样本（中毒样本）：是用于包含后门触发器的训练样本，攻击者在训练过程中使用后门样本将后门嵌入模型中。

• 良性（干净）样本：没有进行任何修改的训练样本。

• 宿主样本：是指被嵌入触发器的中毒样本基于的源干净样本。

• 攻击样本：表示包含后门触发器的恶意测试样本。

• 攻击场景：是指后门攻击可能发生的场景。包括但不限于未受保护的开放渠道、不可靠的数据源和不可靠的训练过程。

• 攻击成功率（Attack Success Rate, ASR）：是指带有后门触发器的测试集被后门感染模型成功误分类为目标标签的概率。

• 攻击后模型预测准确率（After Attack Accuracy, ATA）：表示后门感染模型预测的良性测试样本的准确性。

• 攻击前模型预测准确率（Before Attack Accuracy, BTA）：是使用干净实例进行

训练的无后门的良性模型预测干净测试集的准确性。

• $|P_{BA}| = |BTA - ATA|$：度量后门感染模型的准确性。即后门感染模型与良性模型在遇到良性样本时对其的预测准确率比较。

其中本文使用的评价指标包括攻击成功率 ASR、攻击后模型预测准确率 ATA和 $|P_{BA}|$。满足准确性和有效性目标的合格后门攻击应该有较高的ASR和ATA，但$|P_{BA}|$较低。即后门感染模型在遇到带有触发器的中毒样本时应尽量多地将其成功分类为由攻击者指定的目标标签（高ASR），同时当其遇到良性样本时也应尽量多地预测其实际的源标签（高ATA）。因要求后门感染模型在遇到良性样本时表现良好，故后门感染模型与良性模型在遇到良性样本时对其的预测准确率相比，差距应当尽量小（低$|P_{BA}|$）。

§2.1.2 后门威胁模型实体

本文主要研究图像分类任务中的有毒标签后门攻击问题。在后门感染模型的生命周期中有三个主要的实体：攻击者、模型开发人员和防御者。

（1）攻击者。在本文的威胁模型中，均遵循 BadNets[16]的假设，即攻击者可以访问和操作训练数据，但无法访问受害者模型的参数、结构和训练过程。例如，攻击者可能是训练数据的提供者，通过在干净的实例上嵌入一个自定义的触发器，并将它们的标签更改为目标标签，从而毒害一小部分训练数据。在推理预测阶段，攻击者可以使用包含触发器的攻击样本查询后门感染模型，但是攻击者不能操纵推理过程。

（2）开发人员。开发者可以成为训练后门感染模型的第三方平台。他拥有强大的资源，通常非常专注于训练过程。开发者会仔细选择网络架构、超参数以及训练策略，以获得性能良好的模型。由于在训练过程中涉及的数据量巨大，如果没有明显的异常，例如某些数据有明显的修改痕迹，开发者就不会仔细检查数据的合法性。然而，如果训练过的模型在验证数据集上表现不佳，开发者将拒绝它。

（3）防御者。在模型训练后，防御者可以采取措施，包括检测和缓解后门，如我们在第 1.2.3 节中介绍的，以禁用可疑模型的可能后门。在现实场景中，防御者可以访问可疑模型，并拥有部分源训练数据。他还可以微调或改变模型结构。例如，他可以使用可用的源训练数据对模型进行微调或精细剪枝防御，以去除后门或通过过滤数据抑制后门行为。

§2.1.3 后门攻击的目标

攻击者试图通过数据中毒向受害者模型注入后门。理想的后门攻击应该具有良好

的攻击效果和攻击鲁棒性。良好的攻击效果是攻击成功的基本要求，通常会考虑准确性和攻击有效性。攻击鲁棒性是后门攻击的更高要求，同时也攻击满足隐秘性要求。具体来说，应该拥有以下属性。

（1）准确性。当模型被感染后门后，模型中后门的存在不应降低模型在良性实例上的预测准确性。可以合理地假设，在验证数据上的性能明显低于开发人员预期的后门感染模型将被拒绝部署。

（2）攻击有效性。后门可以很容易地被攻击者制作的特定触发器激活。也就是说，当后门感染模型接收到包含触发器的输入时，模型将极有可能返回由攻击者指定的目标标签，而不管这些输入的实际真实标签是什么。

（3）隐秘性。它要求触发器必须是自然的，通过肉眼或基于梯度的类激活图防御的探测器很难将有毒数据与自然输入区分开来，并且它们的触发器占比区域应该尽量保持在最小或是不突兀。否则，训练数据中的异常将被模型开发人员检测到，并且在训练模型之前，有毒的数据将被清除。

（4）鲁棒性。在一些常用的防御下，如微调或精细剪枝下，后门仍然有效且不易被筛除。

§2.2 卷积神经网络模型介绍

本文主要使用的图像分类网络模型为 VGG16 和 ResNet18，以下为对这两类网络基于的卷积神经网络知识以及对这两个网络的主要介绍。

§2.2.1 卷积神经网络

人类在辨别图像样本时常常会自发寻找样本中一些局部的显著特征而后对图像作出判断，比如通过眼圈判断大熊猫。而神经网络通过矩阵图对上述特征进行描述 ，且对于矩阵图内部的特征提取则可以通过本小节介绍的卷积网络中的卷积核提取。

卷积神经网络[44][45][46]是近年来发展起来的一种高效的识别网络，与传统神经网络相比是功能、形式均不尽相同的层级结构。通过将卷积神经网络的主要结构即卷积层、池化层在不同的循环中交替出现，可以得到具有不同的层次结构的不同卷积网络模型。该网络主要层详解如下。

（1）卷积层。卷积层是卷积神经网络的核心部分，当使用图像数据集训练该类模型时，特征提取有关操作都是交由卷积层完成的。当卷积神经网络接收图像样本输入时，和传统网络不同的是会将其识别为一个由像素值组成的多通道矩阵图，而传统神经网络则是将样本图像识别为向量，故本节介绍的网络和传统神经网络相比的优势之一是不会导致样本图像失去空间化的信息。该层网络通过不同的卷积核矩阵在输入

矩阵图上移动和进行矩阵运算后将图像样本中不同的特征提取出来。在卷积层中，多个核共享相同权值的总体空间位置形成特征图，一个核对应下一层的一个特征图。卷积层可以满足下列公式：

$$y_j^l = f(\sum_i x_i^{l-1} * w_{ij}^l + b_j^l) \tag{2-1}$$

其中 x_i^{l-1} 为 $(i-1)$ 层特征图的第 $(i-1)$ 个通道，$*$为卷积运算。w_{ij}^l 和 b_j^l 分别表示对应层的第 j个核和偏置。符号 y_j^l 为第 l 层特征映射的第 j 个通道，$f(\cdot)$ 表示应用非线性变换的激活函数，如校正线性单元（ReLU）和 sigmoid 函数。

（2）池化层。池化层主要处理卷积层输出的维度较大的特征，可以大大减小矩阵的尺寸，方便进行后续处理。故通常在每个卷积层之后添加池化层，以减少输入样本大小和计算量。随着计算量的减少也预防了过拟合。池化操作可描述如下:

$$Z_j^l = p_{n\in R}(y_j^l(o)) \tag{2-2}$$

其中 $p(\cdot)$ 为池化操作（如常见的平均池化、最大池化），R 为池化区域，o 为位置坐标，Z_j^l为第 l 层特征图第 j 个通道的池化输出。

（3）全连接层。全连接层的功能是整合上层的局部特征，每个节点都与上层的节点连接。它通过权值矩阵形成一个新的图，成为全局特征。全连接层可以将特征矩阵转换为单个值，这样可以有效减少数据量，同时让分类决策不再大程度被特征所处的位置影响。

图 2-1 典型的卷积神经网络基本结构

§ 2.2.2 VGG16

VGG 网络（Visual Geometry Group Network，VGG）[47]是一种深度卷积神经网络在图像分类任务上具有良好的性能，在 2014 年提出。虽然当网络模型中卷积层选择用大卷积核时可以获取大的特征图，全局特征可以被更好地提取。但是使用更小的卷积核可以加深网络的深度，减少计算量的同时增加了更多了非线性表达，让网络的语义表达能力增强。VGG 网络便采取了以上的方式实施卷积核的选取，通过多个小卷积核代替了大卷积核。

本文所使用到的网络结构时 VGG16，网络结构如图 2-2 所示，以下为该网络在本文中训练后门感染模型的流程举例，将 2.3 节中输入大小被统一的数据集输入卷积层。

由卷积核对良性样本或中毒样本进行特征采集后进入池化层。循环 4 轮后良性样本或中毒样本特征信息即可进入全连接层。VGG16 卷积核大小均为 3×3，池化核是 2×2 大小，两层均采用“same”填充方式。每层卷积后跟 ReLu 激活函数，最后全连接层有 3 层，本文对两个数据集各抽取 12 类分别训练，故 VGG16 最后一层全连接层输出分类需设置为 12。

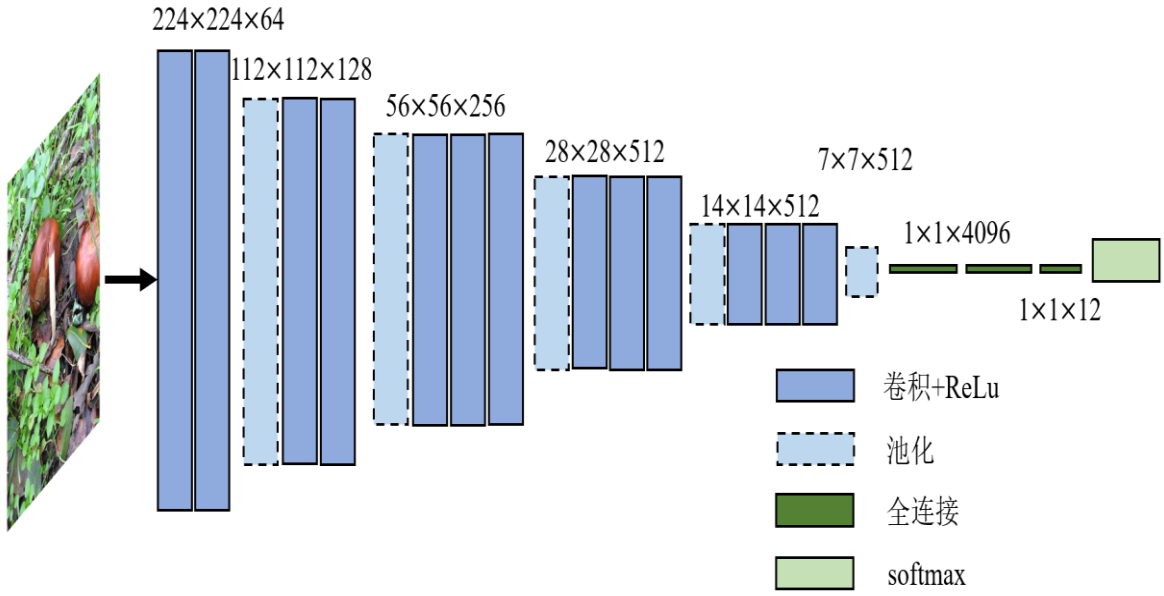


图 2-2 VGG16 卷积神经网络基本结构

§ 2.2.3 ResNet

（1）ResNet[48]残差网络。在人们的理想假设中，随着神经网络的层数堆叠地越深，模型处理问题的能力也会加强且模型的性能会越来越好。但是事实上并非如此，研究表明随着神经网络层数逐渐变深，网络模型的性能曲线虽然是先呈现上升趋势，但是随着网络层数被堆叠地越来越深之后，网络性能反而下降即产生了网络退化的现象。有一部分原因是 ReLu 函数将很多信息置零造成信息不可逆的损失从而导致具有传统结构的网络难以拟合恒等映射，而 ResNet 网络可以通过引入残差模块弥补这类信息损失，解决网络层数变深后出现的网络退化现象。

（2）ResNet18。ResNet18 和 ResNet34 使用 BasicBlock 作为基本单元，具体结构见图 2-3。而层数更深的几个网络结构的基本单元则是使用了 BottlenetBlock，该类型的单元为了减少操作则包含两个 1×1 卷积操作。本文攻击使用的网络结构为 ResNet18，结构如表 2-1 第三列。

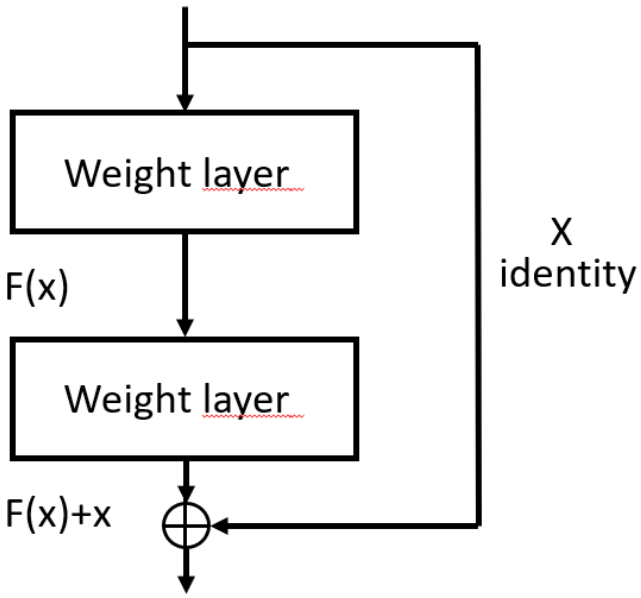


图 2-3 残差模块基本结构

表 2-1 ResNet 结构

层名	输出大小	18 层	34 层	50 层	101 层
Conv1	112×112	7×7，64，步长 2 3×3最大池化层，步长 2			
Conv2_x	56×56	$\begin{bmatrix} 3 \times 3.64 \\ 3 \times 3.64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3.64 \\ 3 \times 3.64 \end{bmatrix} \times 3$	$\begin{matrix} 1 \times 1.64 \\ [\,3 \times 3.64\,] \times 3 \\ 1 \times 1.256 \end{matrix}$	$\begin{matrix} 1 \times 1.64 \\ [\,3 \times 3.64\,] \times 3 \\ 1 \times 1.256 \end{matrix}$
Conv3_x	28×28	$\begin{bmatrix} 3 \times 3.128 \\ 3 \times 3.128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3.128 \\ 3 \times 3.128 \end{bmatrix} \times 4$	$\begin{matrix} 1 \times 1.128 \\ [3 \times 3.128] \times 4 \\ 1 \times 1.512 \end{matrix}$	$\begin{matrix} 1 \times 1.128 \\ [3 \times 3.128] \times 4 \\ 1 \times 1.512 \end{matrix}$
Conv4_x	14×14	$\begin{bmatrix} 3 \times 3.256 \\ 3 \times 3.256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3.256 \\ 3 \times 3.256 \end{bmatrix} \times 6$	$\begin{matrix} 1 \times 1.256 \\ [\,3 \times 3.256\,] \times 6 \\ 1 \times 1.1024 \end{matrix}$	$\begin{matrix} 1 \times 1.256 \\ [\,3 \times 3.256\,] \times 23 \\ 1 \times 1.1024 \end{matrix}$
Conv5_x	7×7	$\begin{bmatrix} 3 \times 3.512 \\ 3 \times 3.512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3.512 \\ 3 \times 3.512 \end{bmatrix} \times 3$	$\begin{matrix} 1 \times 1.512 \\ [\,3 \times 3.512\,] \times 3 \\ 1 \times 1.2048 \end{matrix}$	$\begin{matrix} 1 \times 1.512 \\ [\,3 \times 3.512\,] \times 3 \\ 1 \times 1.2048 \end{matrix}$
—	1×1	平均池化层，1000-d 全连接层，softmax			
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9

§2.3 实验数据集

§2.3.1 ImageNet 数据集

本文第三、四章实验采用的数据集之一 ImageNet[49]数据集是应用于深度学习图像领域的大规模图像数据集，该数据集对许多关于图像的深度学习研究作出了不可忽视的贡献，已经跻身重要基准数据集前列参与众多深度领域里的图像算法研究。ImageNet 数据集涵盖 21841 个类别，总计 14197122 张图像。因 ImageNet 具有类别涵盖广以及数据量庞大等特点，时常被用于验证模型和算法的泛化性能，且该类数据集因为涵盖类别数量众多所以会包含很多相似的数据类别，对分类神经网络模型来说的很好的考验，可以促进其发展，所以是图像分类任务中被用来验证模型泛化性能最常用的数据集之一。

部分样例如图 2-4 所示。该数据集中的图像尺寸大小不一，在图像数据预处理时，本文将所有图像均统一设置为 224×224。

图 2-4 ImageNet 数据集

§2.3.2 GTSRB 数据集

本文第三、四章实验采用的数据集之一是德国交通标志数据集（German Traffic Sign Recognition benchmark，GTSRB）[50]，GTSRB 是 IJCNN2011 举办的一个图像分类挑战赛标准数据集，在图像分类领域具有代表性。GTSRB 共包含 43 类交通标志，训练样本 39209 张，测试样本 12630 张，每幅图像只包含一个标志。这些图像是在包含不同干扰和噪音的自然环境中拍摄的。其中含有大量的低分辨率图像，交通标志牌由不同程度的遮挡、模糊、倾斜等情况。部分交通标志图像如图 2-5 所示，图 2-5 是从全部 43 类图像中的部分类中挑选了一些图像。

图 2-5 GTSRB 数据集

GTSRB 数据集中的图像尺寸大小不一，在图像数据预处理时，本文将分辨率过低的图像筛除，并对剩余图像进行了裁剪等数据增强，所有图像均设置为 224×224。

§2.4 本章小结

本章节首先说明了后门攻击的威胁模型与攻击目标、后门攻击过程中的实体，以及后门攻击的专业术语如评价指标等；之后介绍了本文中使用到的两个卷积神经网络：VGG16 和 Resnet18；在介绍完基础的图像分类神经网络相关知识后，讲解了在本文的后门攻击任务中使用到的两类数据集。通过本章节的介绍，可以对后门攻击任务有更加清晰的认识。

第三章 基于雨滴触发器的后门攻击方法研究

在本节中提出了基于雨滴触发器的后门攻击方法 RDBA（Raindrop Backdoor Attack，RDBA），解决了目前大多数后门攻击在固定位置使用同一个触发器处理不同的干净数据，或者嵌入触发器与宿主样本内容相关性差以及后门样本仅由触发器与干净样本简单叠加生成的问题。本节首先介绍研究动机，接着详细介绍后门攻击整体流程，在触发器生成阶段，通过随机噪声与 α 值生成初步噪声图，再将对角矩阵与旋转矩阵通过仿射变换得到对角核，并使用进行高斯模糊后的对角核对初步生成的噪声图滤波，使得触发器从最初的随机噪声图到有宽度、长度、运动模糊的雨滴图。然后，在后门嵌入阶段，中毒样本和干净样本一同用于训练 DNN 学习从触发器到目标标签的映射。最后则通过实验验证 RDBA 方法的准确性、攻击有效性、隐秘性和鲁棒性。

§3.1 研究动机

由于后门攻击试图影响特定输入的模型预测，已成为深度神经网络模型的严重威胁。然而目前，最流行有效的后门触发器往往是在固定位置使用同一个触发器处理不同的干净数据，或嵌入触发器时不考虑宿主数据的内容导致与宿主样本内容相关性差，此外，中毒样本可能是通过简单地将触发器与良性宿主样本直接叠加而生成。以上触发器生成的后门样本不可避免地存在异常分布，不能自然地将后门嵌入模型，容易引起模型开发者怀疑，甚至这些显著的触发器容易被人工直接筛除，且大多数现有的后门攻击都可以很容易地通过普通防御来阻止。

故在本章节中，我们提出了一种用于图像分类任务的隐秘后门攻击:雨滴后门攻击，其触发器模式在内容上与干净样本融合度高，其触发模式是自然的、不易被察觉的。我们使用雨滴作为后门触发器，它们自然地与干净的实例合并，以合成中毒样本，此类中毒样本与实际自然界中的雨滴图相接近，且分散在图像上的雨滴比文献中的触发器更加多样化，对比于其他触发器是固定的、受限的、触发器内容与干净样本不相关的模式，可以自然地将后门嵌入网络，且该触发器更加隐秘且有效。

(a) ImageNet (b) GTSRB

图 3-1 不同后门方法中毒实例

§3.2 模型整体架构

本节的主要工作包含以下几个方面：首先通过随机噪声与 α 值保证雨滴触发器的均匀随机分布以及控制触发器密度，再通过构建好的对角矩阵与旋转矩阵进行仿射变换得到一个对角核，最后对该核进行高斯模糊得到模糊核且使用该模糊核对初步生成的噪声图进行滤波操作，使得触发器完成由从最初的随机噪声图到有宽度、长度、运动模糊的雨滴图的转变。提出的触发器生成算法框架如算法 3-1 所示。

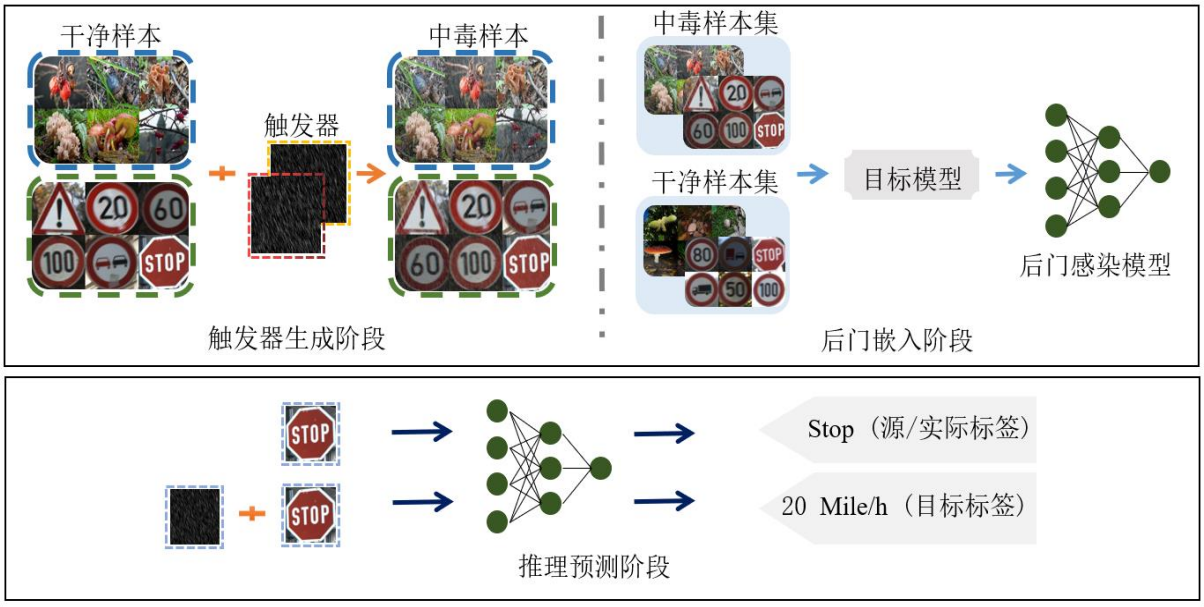


图 3-2 RDBA 流程图

将本章节所提出的基于自然特征雨滴的后门模型 RDBA，其整体攻击流程如图 3-2 所示。在触发器生成阶段，攻击者利用雨滴触发器毒害一小部分干净训练样本，生成后门样本。在后门嵌入阶段，后门样本和干净样本一同被注入 DNN 中用来训练 DNN 得以学习从雨滴触发器到由攻击者选择的目标标签的映射。在预测推理阶段，后门模型用于返回干净测试输入的样本真实标签和用于返回中毒测试输入样本的目标标签。

§3.3 后门攻击具体流程

§3.3.1 模型参数定义

在不失一般性的前提下，本章在一个C类别的图像分类任务上考虑 DNN 后门攻击问题。假设$D=\{(x_i,y_i)\}_{i=1}^{N}$表示来自可信数据源的包含了$N$个样本数据的干净数据集。其中$x_i\in\{0,\dots,255\}^{w\times h\times c}$表示干净样本、$y_i\in\{0,\dots,C-1\}$表示与干净样本相对应的源标签。

我们让$y^t\in\{0,\dots,C-1\}$表示攻击者选择的目标标签并遵循 Wang 等人[51]在文献[51]中的定义，将我们的数据中毒算法$A(\cdot)$定义如下：

$$x_i^t\leftarrow A(x_i,m,\Delta), \tag{3-1}$$

$$x_{jkc}^t=\left(1-m_{jk}\right)\cdot x_{jkc}+m_{jk}\cdot\Delta_{jkc}. \tag{3-2}$$

其中$x_i\in D_1$表示原始良性干净样本，D_1是D的子集，x_i^t为中毒样本，Δ为触发器，m是一个二维矩阵，将其称为掩码，c、w、h分别为图像通道数、图像宽度、图像高度。

后门模型F_B的总体训练集是少数带有后门触发器的中毒样本和剩余的干净样本的组合，其中后门触发器训练样本集$D_{trigger}=\{(x_i^t,y^t)\}_{i=1}^{|D_1|}$，剩下的干净数据集为$D_2=D\backslash D_1$，总的训练集$D_{train}$定义如下:

$$D_{train}=D_2\cup D_{trigger} \tag{3-3}$$

则中毒样本注入率为：$k=\frac{|D_{trigger}|}{|D_{train}|}$。

§3.3.2 生成雨滴触发器

在基于雨滴触发器的后门攻击中，用来毒害干净实例样本的触发器采用自然特征：雨滴，此类自然特征可以与一些室外数据集在内容语义上融合良好。具体的雨滴触发器生成步骤如下：

对于每一个样本 $x \in D_1$，我们首先生成随机噪声：

$$noise \leftarrow \{random(0,256)\}_{i=1}^{\omega \times h} \tag{3-4}$$

为了使最终生成的雨滴触发器 Δ看起来自然且隐秘，我们使用一个 α值对噪声进行预处理以约束雨滴密度，然后用卷积核 K_1进行第一次滤波运算：

$$noise = \begin{cases} 255, & if\ noise > (256-\alpha), \\ 0, & else. \end{cases} \tag{3-5}$$

$$\Delta_1 \leftarrow B(noise, K_1) \tag{3-6}$$

其中 K_1为单通道 3×3 浮点矩阵。为了进一步实现雨滴，需要对初步生成的雨滴触发器进行拉伸和旋转，以模拟不同大小和方向的雨滴，这一步使用构建好的对角矩阵与旋转矩阵进行仿射变换得到的对角核来完成，其中对角矩阵对角线值为 1，其余值为 0，然后利用高斯模糊对该核添加运动模糊后得到模糊核K_2。使用高斯模糊需要构造相应的权值矩阵进行滤波，权值的计算依赖于二维高斯函数。下面是使用的二维高斯函数:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{3-7}$$

雨滴触发器通过应用模糊核K_2的第二次滤波操作来更新，高斯模糊可以使得雨滴触发器看起来具有真实的运动模糊:

$$\Delta \leftarrow B(\Delta, K_2) \tag{3-8}$$

然后，对于 D_1中的所有样本，重复上述步骤，应用公式（3-2）中的算法，得到我们的雨滴触发器样本 $D_{trigger}$。详细的雨滴触发器生成过程在算法 3-1 中进行了总结。

§3.3.3 嵌入后门

使用上述方法生成后门触发器训练样本集 $D_{trigger}$后，攻击者会用它更新训练数据集 D_{train}。模型开发人员使用 D_{train}训练神经网络模型，采用交叉熵损失的标准模型训练过程，即解决以下优化问题:

$$\underset{\theta}{\operatorname{argmin}} -\frac{1}{N} \sum_{x \in D_{train}} \sum_{i=1}^{C} y_i \log\big(p_i(x,\theta)\big) \tag{3-9}$$

式中，y_i为 x的实际真值标签的第i个值，p_i为 F_B的 $Softmax$的第 i个输出，θ为可训

练模型权值。优化式（3-9）可以用 SGD（随机梯度下降）优化器的反向传播来求解。

由于后门训练数据集包含了κ比值有毒的数据样本，模型可以学习从触发器到目标标签的映射，即在训练过程中后门会嵌入到模型中。在推理预测阶段，攻击者可以通过将触发器注入到良性样本中并将其输入到模型中来激活后门行为。

算法 3-1：RDBA 生成雨滴触发器伪代码

算法 3-1：RDBA 生成雨滴触发器伪代码

输入:干净训练样本集$\boldsymbol{D}=\{(\boldsymbol{x_i},\boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$，中毒样本注入率 κ, 触发器密度（雨滴密度）$\boldsymbol{\alpha}$，模糊核$\boldsymbol{K_1}$和$\boldsymbol{K_2}$, 掩码$\boldsymbol{m}$。

输出:后门触发器训练样本集（中毒数据集）$\boldsymbol{D_{trigger}}$

1：将噪声、触发器初始化为 0、后门触发器训练样本集初始化为空数组

2：在干净训练样本集 $\boldsymbol{D}$中随机选取比例κ的数据样本放入 $\boldsymbol{D_1}$

3：**for all** $\boldsymbol{x}\in\boldsymbol{D_1}$ **do**

4：　随机生成范围从 0 到 255 的灰度值，赋值给 $\boldsymbol{noise}$

5：　**if** $\boldsymbol{noise}<(\boldsymbol{256}-\boldsymbol{\alpha})$

6：　　将$\boldsymbol{noise}$置为 $\boldsymbol{0}$，此部分噪声为黑色

7：　**else**

8：　　将$\boldsymbol{noise}$置为 $\boldsymbol{255}$，此部分噪声为白色

9：　使用模糊核 $\boldsymbol{K_1}$对噪声做第一次滤波操作得到初始触发器

10：　使用模糊核 $\boldsymbol{K_2}$对噪声做第二次滤波操作得到雨滴触发器

11：　使用掩码 $\boldsymbol{m}$，将白色雨滴噪声部分叠加在干净样本上，黑色噪声部分则保留宿主样本源内容

12：　将 $\boldsymbol{x^t}$加入后门触发器训练样本集 $\boldsymbol{D_{trigger}}$中

13：**end for**

14：**返回** $\boldsymbol{D_{trigger}}$

§3.4 实验结果分析

为了评估提出的后门方法在攻击效果和攻击鲁棒性等方面的性能，本章节使用两种不同的基准数据集和两种神经网络结构进行了大量的实验。并使用了两种最流行的有毒标签后门方法，Gu 等人[16]提出的 BadNets 和 Chen 等人[28]提出的 Blending，作为本节的参考基准。

§3.4.1 实验参数设置

本章节评估了后门攻击在两个基准数据集上的性能:ImageNet[49]和 GTSRB[50]。为

了简单起见，本节从两个数据集中随机选择一个连续的包含 12 个类别的子集，用于训练和测试，其中第一个类别被定义为目标类。对于 ImageNet 和 GTSRB，所选的子集分别包含 15,592 张图片和 40,520 张图片。将两个子集均按 10:1 的比例分成训练集和测试集，并采用数据增强方法（随机裁剪和旋转）对样本进行处理。这些图像都被调整为 244×244×3。

所有对这两个数据集的训练都是在 ResNet18[32]和 VGG16[33]上进行的，每个数据集分别在两个模型上训练了 200 个轮次，ImageNet 数据集大概需要 23 个小时完成训练、GTSRB 大约需要 13 个小时完成训练。中毒样本注入率默认为 $\kappa = 0.09$。对于 RDBA，雨滴触发器密度默认为 $\alpha = 6$。在训练阶段使用 SGD[52]优化器，初始学习速率设置为 0.01，训练批量大小和最大迭代周期分别设置为 32 和 200。

我们使用的评价指标包括ASR（攻击成功率）、ATA（攻击后准确率）和P_{BA}。ASR 是指带有后门触发器的攻击样本被有毒模型误分类为目标标签的概率。ATA是指有毒模型在干净测试集上的性能。$P_{BA} = BTA - ATA$度量后门感染模型的预测准确性，其中BTA（攻击精度之前）是使用训练集均为干净样本进行训练的良性模型对于干净测试集的预测准确性。满足准确性和攻击有效性目标的合格后门攻击应该有较高的 ASR 和ATA，且$|P_{BA}|$较低。

§3.4.2 实验结果分析

本节从四个角度讨论了 RDBA 的性能，分别为：准确性、攻击有效性、隐秘性、鲁棒性。

表 3-1 不同后门方法性能其中x/y表示平均指标ATA/ASR，最好的结果用粗体表示

Dataset	Model	$\boldsymbol{BTA}$	Blending[28]	BadNets[16]	RDBA
GTSRB	ResNet18	93.87	90.04/99.80	93.05/99.13	**93.52**/**99.94**
	VGG16	92.31	92.83/99.97	**93.22**/97.39	92.86/**100**
ImageNet	ResNet18	87.30	85.12/**99.32**	84.43/97.24	**86.70**/99.25
	VGG16	86.90	85.34/**99.46**	84.13/92.05	**87.18**/99.19

（1）准确性：它的目的是验证当后门感染模型在测试干净数据时其预测准确率是否会因为后门而受到影响。作为比较，使用实验设置中两种网络结构和干净训练数据集训练得到相应的良性模型，其模型性能如表 3-1 第三列所示。同时，使用 Blending[28]和 BadNets 方法训练得到相应后门模型，精度结果如表 3-1 所示。比较 RDBA 方法的 ATA和 BTA值，很明显由 RDBA 方法的得到的后门感染模型对准确性性能没有负面影响，且在 VGG16 上训练的模型对于干净数据集的预测准确度甚至有轻微的提高。但是，从表 3-1 第四和第五列可以看出，由 Blending 和 BadNets 方法训练得到的后门

感染模型在准确性指标上表现得不是很好。例如，在 Blending 过程中，使用 GTSRB 训练的 ResNet18 模型，其 ATA相较于良性模型下降了 3.83%；而在 BadNets 中，用 ImageNet 训练的 ResNet18 模型，其 ATA则较于良性模型下降了 2.87%。

除此之外，本节还研究了不同雨滴触发器密度 α 和不同中毒样本注入率 κ对后门感染模型在准确性指标上的影响，实验结果分别见表 3-2 和图 3-3。从表 3-2 可以看出，$|P_{BA}|$值普遍较小，最大值也仅为为 0.56%，这说明了由 RDBA 训练生成的后门模型与良性模型在遇到干净测试集时表现差距很小，一般可以认为可以忽略不计。

从图 3-3 可以看出，在使用 GTSRB 数据集训练同时后门样本注入率为 0.08 时、及在使用 ImageNet 数据集训练同时后门样本注入率为 0.04 和 0.08 时 ATA略有下降，但随着中毒样本注入率的增加，ATA总体保持相对稳定。总之，RDBA 方法实现了高准确性，能够保证后门嵌入的同时不影响后门感染模型对于干净输入的预测准确性。

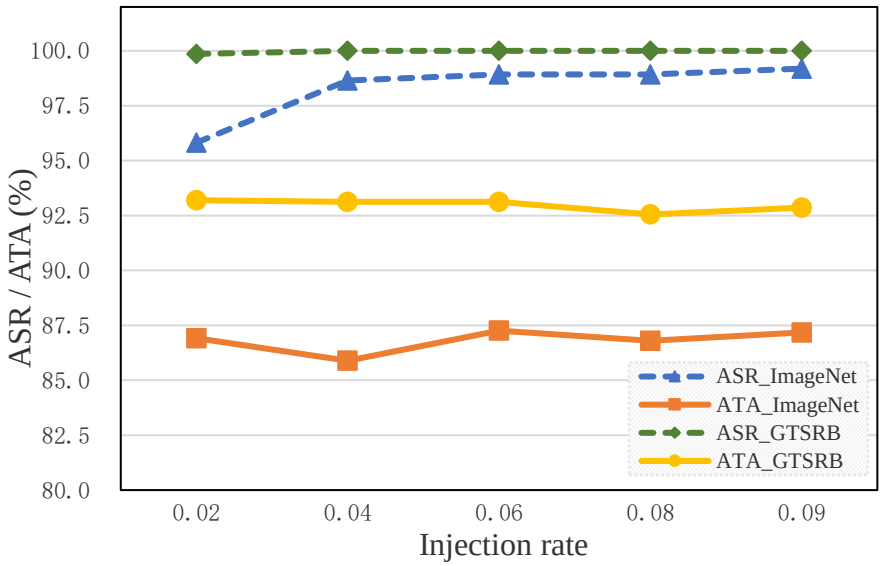


图 3-3 不同后门样本注入率对 RDBA 的影响

（2）攻击有效性：攻击有效性的目的是量化包含特定触发器的攻击样本激活目标标签的可能性。从表 3-1 可以看出，几乎所有后门攻击方法的 ASR都很高。对于 RDBA 方法和 Blending 方法来说，它们的 ASR接近 100%。而对于使用 ImageNet 在 VGG16 上训练的 BadNets 来说，ASR 约为 92%，这意味着基于 BadNets 的后门攻击有效性相对于 RDBA 和 Blending 方法来说是较差的。

表 3-2 进一步显示，RDBA 的ASR随着密度 α 的增加而增加。当 α为 0.5 时，其 ASR为 96.42%，虽然相对较低，但仍优于仅为 92%的 BadNets。当 α增加到 1 时，ASR接近 99%，而之后随着触发器密度增加，后门攻击成功率均能稳定在 99%以上。这证明了雨滴触发器的有效性，更说明即使雨滴触发器在低密度下，RDBA 的后门攻击成功率也很高。

表 3-2 不同密度的雨滴后门触发器性能

α	ATA	ASR	$\lvert P_{BA}\rvert$
0.5	87.15	96.42	0.25
1	86.34	98.85	0.56
2	87.05	99.12	0.15
3	86.86	99.32	0.04
4	87.03	99.39	0.13
5	86.97	99.52	0.07
6	87.18	99.10	0.28

图 3-3 也显示了相似的趋势，即对于基于 RDBA 方法的后门攻击来说，ASR 随着后门样本注入率的增加而增加。虽然当注入率为 0.02 时，在 ImageNet 数据集上训练的 RDBA 的 ASR约为 95%，低于在 GTSRB 上训练的接近 100%的 ASR。但这主要是因为 ImageNet 数据集的特征表示更加复杂导致其分类难度高于 GTSRB 数据集。一旦当中毒样本注入率增加到 0.04 时，ASR 便可高达 99%左右。之后，随着注射率不断增加，ASR值则一直稳定在 99%~100%之间。综上所述，触发器密度的设置对后门感染模型的效果没有显著的影响，本节所提出的 RDBA 方法即使在较小触发器密度或低中毒样本注入率的情况下也能够达到较高的攻击成功率，故基于 RDBA 的后门攻击方法攻击有效性高。

（3）隐秘性：隐秘性是用来衡量这些中毒样本引起开发者怀疑的可能性有多大。从直观上看，嵌入后门触发器后图像越自然、中毒样本注入率越小，被嵌入触发器的数据就越隐秘，模型开发人员就越不可能注意到它们。

(a) α = 0.5

(b) α = 3

(c) α = 6

图 3-4 不同雨滴触发器密度的中毒实例

图 3-4 显示了不同触发器密度生成的中毒样本实例。从图中可以观察到，即使对干净样本的扰动随着触发器密度的增加而增加，修改后的图像在肉眼看来还是很自然的。图 3-1 显示了不同方法中使用的后门实例。很明显，由 Blending 和 BadNets 创建的后门实例有明显的人工合成的痕迹，且与宿主样本的内容相关性较差。相比之下，使用 RDBA 创建的中毒样本看起来更自然，宿主样本的内容不受影响，而正如在章节 2.1.2 中所说对于一个预算有限的模型开发人员来说，由于在训练过程中涉及的数据量大，如果没有明显的异常如某些数据有明显的修改痕迹，开发者就不会仔细检查数据的合法性。即如果训练数据集没有明显的异常，开发人员就不会意识到后门问题。从这个意义上说，雨滴被用作后门攻击的触发器隐秘性更强。另一方面，正如我们在（1）、（2）中分析的那样，RDBA 可以在相对较低的中毒样本注入率下实现高准确性和攻击有效性。例如，如图 3-3 所示，在 ImageNet 上训练的 RDBA 在中毒样本注入率为 0.04 时达到近 99%，而在 GTSRB 上几乎达到 100%。因此，可以得出 RDBA 满足隐秘性标准的结论。

（4）鲁棒性：鲁棒性评估的目的是衡量后门方法是否能够承受和抵抗后门防御。在本节中，本章节中主要关注的几类后门防御是 1）微调防御、2）精细剪枝防御和 3）基于梯度的类激活图防御。

1）微调防御（Fine-tuning）。本章节评估了 Blending、BadNets 和 RDBA 在抵御微调防御方面的效果。通过这三种后门攻击方法和 ImageNet 数据集对后门模型进行预训练。然后使用 10%的干净 ImageNet 数据集对以上后门感染模型进行 10 个周期的微调，学习率设置为 0.001，如图 3-5 所示。可见，BadNets 的 ASR在仅经过 2 个周期的微调后就显著下降，且随着微调周期逐渐增大，ASR 呈持续下降趋势。最后，经过 10 个时期的微调，BadNets 的 ASR从 92.05%下降至近 30%。相反，随着微调周期的增加，基于 Blending 和 RDBA 方法的 ASR在基本上不受影响。基于 RDBA 方法的 ASR在微调周期为 8 时下降了 2.18%，但很快恢复，最终维持在 97%左右。总的来说，基于 RDBA 方法的 ASR性能可以与 Blending 相媲美，而且两者在抵御微调防御方面都优于 BadNets 方法。这可能是因为 BadNets 的触发器模式过于简单，且触发器模式只关注图像的一小部分，导致后门感染模型中预测干净测试样本的神经元很少与预测触发器的神经元重叠。因此，BadNets 更容易受到微调防御的影响。

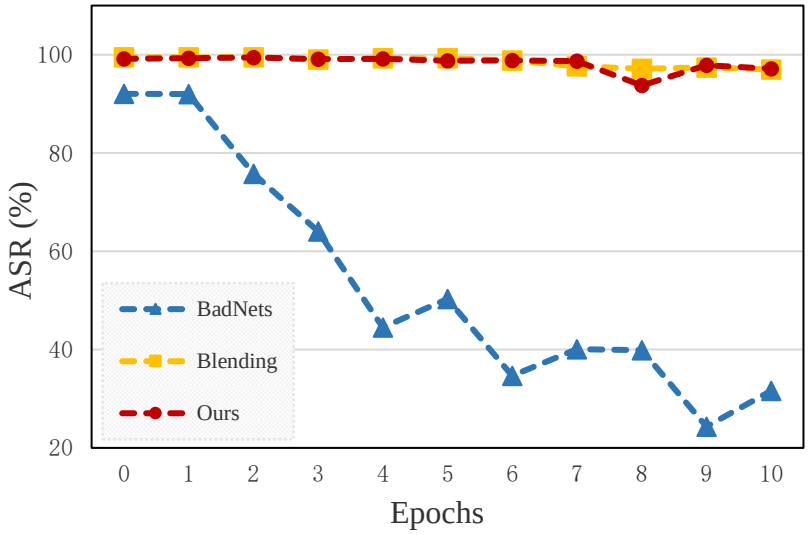


图 3-5 经过微调防御后，不同方法后门模型的攻击成功率 ASR

此外，本节还研究了 RDBA 方法中的雨滴触发器密度对抵御微调防御的影响，如图 3-6 所示，利用不同密度 α 的中毒样本训练后门模型。从图中可以看出，密度为 0.5 和 1 的后门模型 ASR随着微调周期增加而显著下降，最后，经过 10 个微调周期，该值降至近 90%。这是因为当密度较低时，攻击样本与干净样本的相似度较高，使得 DNN 模型无法学习准确区分触发器特征。但同时，当触发器密度为 2 或大于 2 的后门模型 ASR则几乎不再受微调的影响而产生大幅降低，ASR较为稳定地保持在 100% 左右。结果表明，在密度≥2 的情况下，该方法可以很好地保持后门行为。值得一提的是，正如本文之前讨论过的，在这样的设置中，后门的鲁棒性得到了很好的维持。

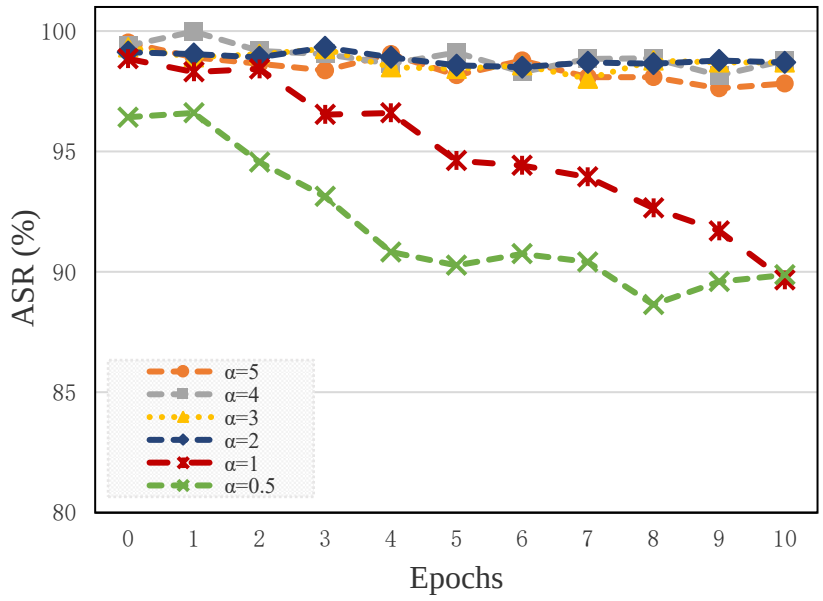


图 3-6 经过微调防御后，不同密度 α的后门模型的攻击成功率 ASR

2）精细剪枝防御（Fine-pruning）。通过设计如下的实验来评估 RDBA 方法对这种防御的后门行为的鲁棒性，在 ImageNet 和 GTSRB 数据集上使用 VGG16 对后门模型进行训练。本节根据大小对权重进行排序，然后将最小的 p%设为 0 来修剪模型。然后，使用 10%的干净数据对经过修剪的模型进行微调。实验结果如图 3-7 所示。从图中我们可以看出，当修剪 20%的神经元时，六种后门感染模型的 ASR都保持得很好，但当修剪的神经元比例超过 20%时，ASR下降明显。当剪枝率为达到 40%时，Blending 方法的 ASR下降最严重，尤其是使用该方法在 GTSRB 上训练的模型，其 ASR从近 100%下降到近 40%。对于在 ImageNet 上使用 Blending 和 BadNets 方法进行后门嵌入的模型，它们的 ASR下降了约 20%。当修剪率达到 50%时，其 ASR下降到 60%左右。与 Blending 和 BadNets 相比，RDBA 不易受到精细剪枝防御的影响。在 ImageNet 和 GTSRB 数据集上，即使该后门感染模型有 50%的神经元被修剪，RDBA 后门模型的 ASR值都在 80%以上。总的来说，提出的 RDBA 在保持精细剪枝防御后的后门攻击成功率方面优于 Blending 和 BadNets。

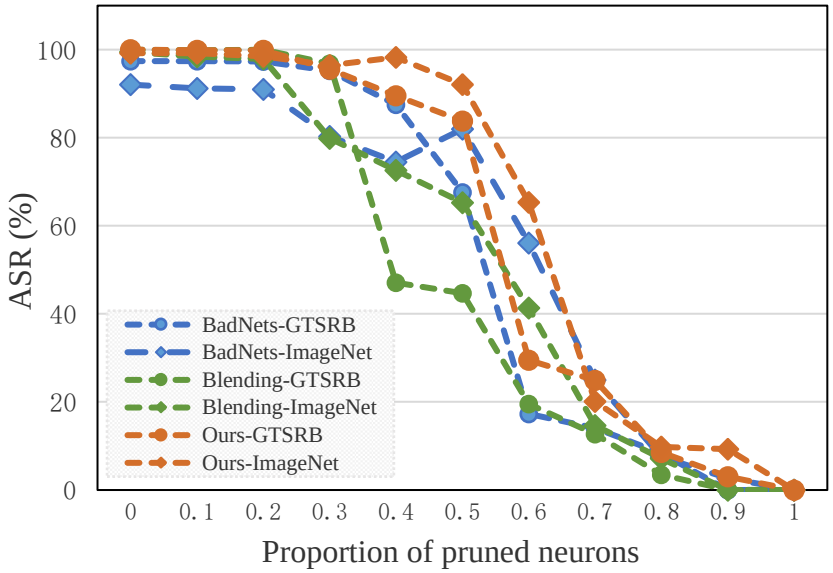
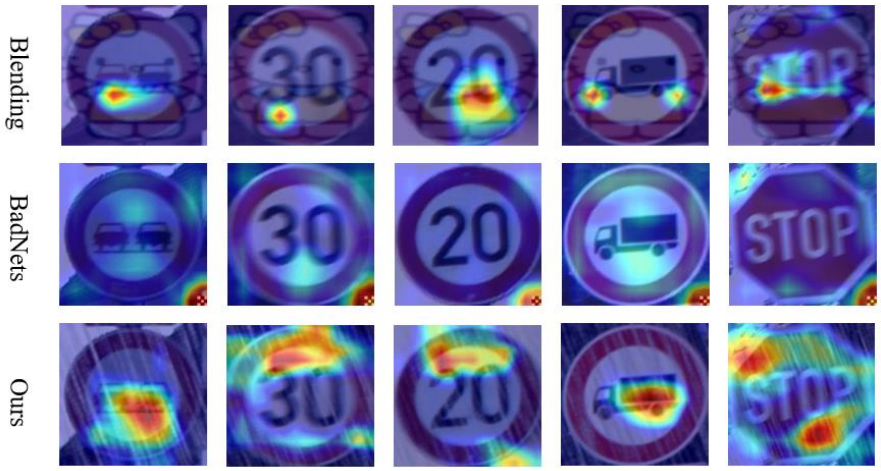


图 3-7 经过精细剪枝防御后，不同方法后门模型的攻击成功率 ASR

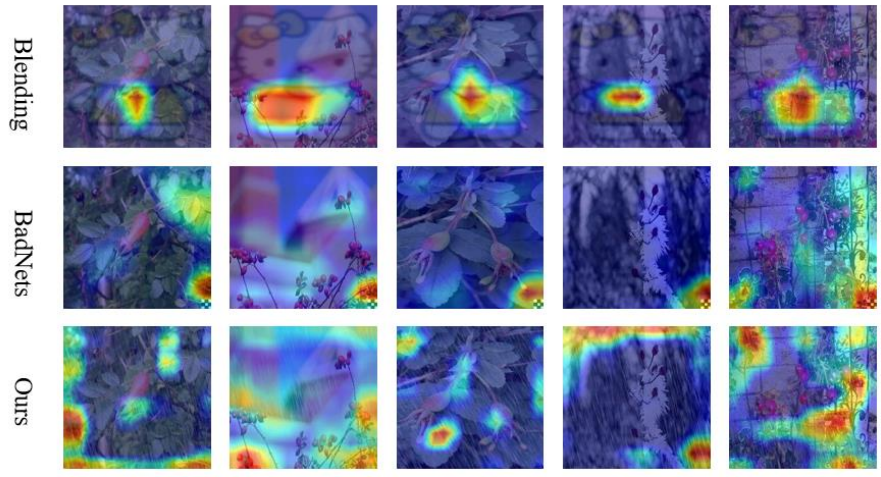
3）基于梯度的类激活图防御（Grad-CAM）。如第二章所述，基于梯度的类激活图防御方法的有效性高度依赖于该方法是否可以精准定位恶意显著区域，即其对于触发器所在区域的定位精度。为了评估本章节的后门攻击方法对于这类防御的抵抗性能，在 VGG16 上对给定的两类数据集，通过 Grad-CAM 方法生成中毒样本实例的类激活图，结果如图 3-8 所示。从图 3-8（a）和（b）的第二行可以看出，由 BadNets 方法所获得的后门样本，其类激活图定位的区域主要集中在特定的显著区域，即中毒样本的右下角，而这一区域正是其触发器嵌入的位置。类似于这种普遍且非常集中的显著区域很可能被开发者识别为恶意的显著区域，如果开发者在使用包含该类中毒样本的训练集时，选择将这一部分样本的激活内容直接截除或者替换补齐为其他内容，对该后

门方法的打击将会是巨大的，甚至可能导致其后门行为彻底失效。对于由 Blending 方法训练得到的中毒样本所生成的类激活图，如图 3-8（a）和（b）第一行所示，一方面，它们的显著区域则并不像前面所提到的 BadNets 方法那样聚集在图像某一部分的固定区域上。但另一方面，由该方法得到的后门样本的类激活图，其高亮的显著区域的分布保持了一定的规律性，大部分集中在图像的下半部分的中间。

与以上两种后门方法相比，由 RDBA 生成的中毒样本的显著区域则更为分离地散布在图像中，如图 3-8（a）和（b）第三行所示，是呈随机分布的。以上结果主要是因为由 RDBA 生成的不同有毒图像包含不完全相同的触发器，且触发器在图像中呈现出均匀分布，而由 Blending 和 BadNets 生成的有毒图像与 RDBA 相比较具有固定的、有迹可循的触发器模式。而且由于使用 RDBA 方法生成的中毒样本所得到的类激活图显著区域分散程度高，开发者更难将此类触发器完全从样本中剥除，开发者如若强行这样做的结果可能会影响剥离触发器后的图像的样本特征表达。综上所述，RDBA 生成的触发器更难区分且更难移除，对基于梯度的类激活图防御更有抵抗力。



(a) GTSRB



(b) ImageNet

图 3-8 不同后门模型实例的 Grad-CAM

§3.5 本章小结

在本章节中，说明了大多数现有的后门攻击遵循所有良性样本共享同样的固定触发器模式，或者嵌入触发器与宿主样本内容相关性差的问题。这不仅使后门样本由于其不自然的外观很容易被模型开发人员怀疑，而且还允许当前的后门防御轻松地阻止后门攻击行为。基于此，我们提出了基于自然雨滴现象的 RDBA 攻击，与现有的后门触发器相比，RDBA 攻击中雨滴触发器均匀地分布在图像上，不遵循所有良性样本共享同样的固定触发器模式，更加隐秘，可以规避数据过滤。这使得 RDBA 对现有的后门防御更具抵抗力。通过大量的实验，验证了 RDBA 在攻击不同模型时在准确性、攻击有效性、隐秘性和鲁棒性方面的攻击效果。

第四章 基于图像隐写触发器的后门攻击方法研究

在本章节介绍了提出的基于图像隐写触发器的后门攻击方法 ISBA。解决了大多数现有的后门攻击方法其触发器模式受限或触发器内容与其宿主样本内容不相关联的问题，以及在第三章节提出的基于雨滴触发器的后门攻击方法不能兼容所有类型的数据集的问题。本节首先在触发器生成阶段，对随机选中的干净样本利用二维离散傅里叶变换实现空间域到频域的转换操作，接着将触发器与以上频域样本结合，再通过逆二维离散傅里叶变换将频域样本转换回空间域得到后门攻击样本。然后，在后门嵌入阶段，中毒样本和干净样本一同被注入 DNN 中用来训练 DNN 学习从隐写触发器到目标标签的映射。在推理阶段，后门感染模型用于返回干净测试输入的样本真实标签和用于返回中毒测试输入的目标标签。最后通过实验验证了 ISBA 各性能。

§4.1 研究动机

由于用于将后门植入受害者模型的有毒样本中后门触发器往往内容大多与干净样本原本的内容相关性差，导致这些显著的触发器容易被人工直接筛除，大多数现有的后门攻击都可以很容易地通过普通防御来阻止。除此之外，在上一节中我们介绍了一种雨滴触发模式，虽然这种触发基于自然的模式，即无论对于开发者或者对于部分防御算法来说，其触发模式是自然的、不易察觉的，但其应用场景局限于室外的场景，但如果触发器是不可见的，则可以较好弥补雨滴触发器的局限性，将应用场景从室外拓展到几乎所有场景的数据。

为了解决以上问题本章节提出了一种基于图像隐写的后门攻击算法，称作 ISBA，该方法生成的触发器是不可见的。首先，通过二维快速傅里叶变换将由攻击者选取的含有目标类别信息的图片作为触发器，将其在频域嵌入少量干净样本中，通过将在频域添加好触发器的中毒样本进行逆傅里叶变换得到最终的中毒样本。再将其注入剩余干净样本中得到攻击样本训练集。最后，将后门样本与干净样本混合共同送入多分类神经网络训练，以得到后门模型。由于这类触发器不可见的特性，将目标信息隐写入干净数据样本中，相较于现有后门工作隐秘性更强且对于不同类型数据集兼容性更强。

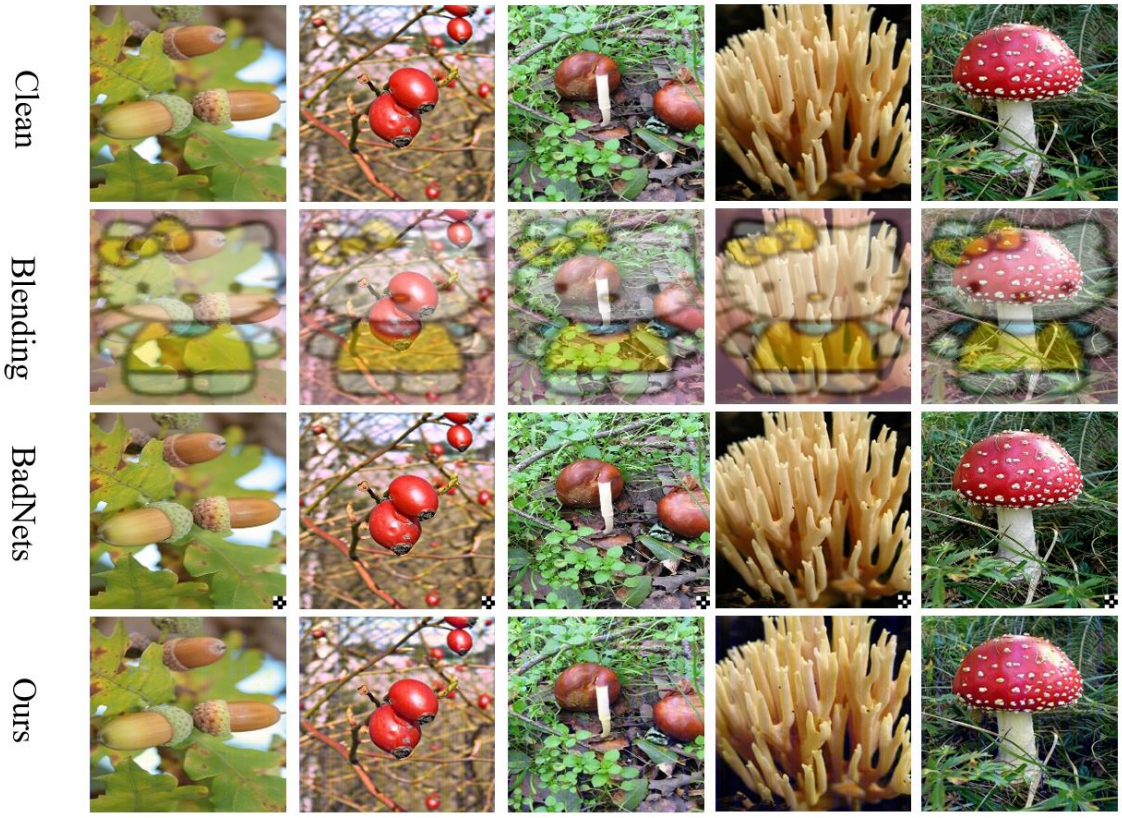


图 4-1 各类后门攻击中毒实例

§ 4.2 模型整体架构

本小节的主要工作包含以下几个方面：首先从干净数据集中取出一小部分样本，对每一个样本利用二维离散傅里叶变换进行空间域到频域的转换操作，然后将由攻击者制作的带有目标类别信息的触发器图像与以上频域样本结合，最后通过逆二维离散傅里叶变换将已被嵌入触发器的频域样本转换回空间域得到后门攻击样本。提出的触发器生成框架如算法 4-1 所示。

本小节所提出的 ISBA 的主要攻击流程如图 4-2 所示。首先，在触发器生成阶段，攻击者利用隐写触发器毒害一小部分干净训练数据，生成中毒样本。然后，在后门嵌入阶段，中毒样本和干净样本一同被注入 DNN 中用来训练 DNN 学习从隐写触发器到由攻击者选择的目标标签的映射。最后，在推理阶段，后门感染模型用于返回干净测试输入的样本真实标签和用于返回中毒测试输入的目标标签。

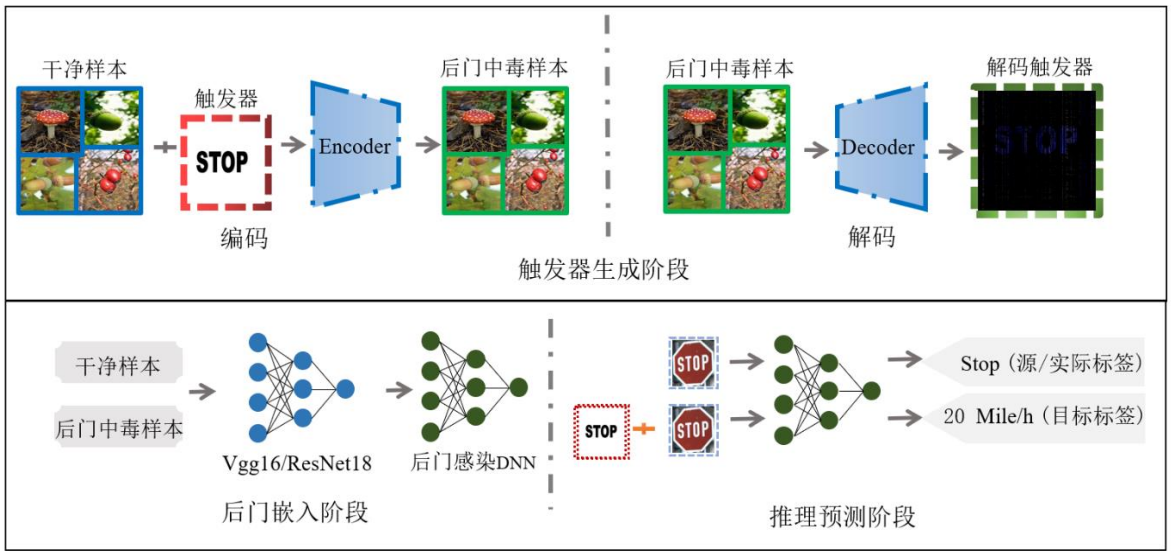


图 4-2 ISBA 后门攻击整体流程

§4.3 后门攻击具体流程

§4.3.1 模型参数定义

本节在一个共有 C个类别的图像分类任务上考虑 DNN 后门攻击问题。假设 $D = \{(f(m,n)_i, y_i)\}_{i=1}^{N}$表示来自可信数据源的包含了 N个样本数据的干净数据集。其中 $f(m,n)_i \in \{0,\dots,255\}^{W\times H\times c}$表示干净样本、$y_i \in \{0,\dots,C-1\}$表示与干净样本相对应的源标签。

让 $y^t \in \{0,\dots,C-1\}$表示由攻击者选择的目标标签。ISBA 整体后门攻击算法$A(\cdot)$定义如下：

$$f(m,n)_i^t \leftarrow A(f(m,n)_i, \Delta) \tag{4-1}$$

其中$f(m,n)_i \in D_1$表示原始良性干净样本，D_1是D的子集，$f(m,n)_i^t$为中毒样本，Δ为触发器， c 表示图像通道数，W、H分别表示图像宽高。

后门模型 F_B的总体训练数据集由两部分组成，第一部分是少数带有后门触发器的中毒样本、第二部分是剩余的干净样本数据集，其中后门触发器训练样本集 $D_{trigger} = \{(f(m,n)_i^t, y^t)\}_{i=1}^{|D_1|}$，剩下的干净数据集为$D_2 = D\backslash D_1$，总的训练集$D_{train}$定义如下:

$$D_{train} = D_2 \cup D_{trigger} \tag{4-2}$$

则后门注入率为：$k = \frac{|D_{trigger}|}{|D_{train}|}$。

§4.3.2 生成图像隐写触发器

在基于图像隐写触发器的后门攻击方法中，用来毒害干净实例样本的触发器采用由攻击者制作的带有目标标签信息的图像，如图4-2中触发器生成阶段的触发器实例。由于此类触发器与干净样本在频域结合后转回到空间域是肉眼不可见的，故而可以与包含室外数据集在内的各类数据集在内容语义上融合良好，对不同种类数据集兼容性更强。其中干净样本由空间域到频域之间的转换通过二维离散傅里叶变换（Two Dimensional Discrete Fourier Transform，2D-DFT）完成。

一维傅里叶变换[53]是将时间域上的信号转变为频率域上的信号，经过傅里叶变换后，对同一事物的看待角度随之改变，可以从频域里发现一些从时域里不易察觉的特征。故而攻击者将干净样本转换到频域可以方便地对样本进行触发器的嵌入。在实际工程应用里使用到的傅里叶变换大都是离散傅里叶变换（Discrete Fourier Transform，DFT）。DFT 是采样的傅立叶变换，因此不包含构成图像的所有频率，而仅包含足够大以完全描述空间域图像的一组采样。频率的数量对应于空间域图像中的像素数量，即，空间域和傅立叶域中的图像具有相同的大小。

一维离散傅里叶公式表示如下：

$$F(u) = \sum_{x=0}^{M-1} f(x) e^{-j2\pi ux/M} \tag{4-3}$$

其中，$u = 0,1,2,\dots,M-1$。

一维离散逆傅里叶变换如下：

$$f(x) = \frac{1}{M} \sum_{u=0}^{M-1} F(u) e^{j2\pi ux/M} \tag{4-4}$$

其中，$x = 0,1,2,\dots,M-1$。

由于灰度图像是由二维的离散的点构成的，故而 2D-DFT 常用于图像处理中，对图像进行 2D-DFT 后可得到其频域图。本节在添加后门触发器时首先将每个 $f_i(m,n) \in D_1$的干净样本通过 2D-DFT 从空间域转换到频域，即 $F_i(u,v)$，便于攻击者将触发器添加到干净样本中。具体的后门触发器由图 4-2 中触发器生成阶段所示。对于大小为 W×H 的干净样本 $f_i(m,n) \in D_1$进行二维离散傅里叶变换公式如下：

$$F(u,v)=\sum_{m=0}^{W-1}\sum_{n=0}^{H-1}f(m,n)e^{-j2\pi(^{um}/_{W}+^{vn}/_{H})} \quad (4\text{-}5)$$

其中$f(m,n)$是干净样本空间域中的图像，大小为$W\times H$，$m\in\{0,\dots,W-1\}$、$n\in\{0,\dots,H-1\}$，公式中指数项是傅立叶空间中每个点$F(u,v)$对应的基函数。该公式可以理解为：每个点$F(u,v)$的值是通过将空间图像与相应的基函数相乘并求和得到的。

将少量干净样本从空域转换到频域后，将触发器图像 Δ 添加到干净样本频域中得到初步的后门样本$F(u,v)'$：

$$F(u,v)'\leftarrow E(F(u,v),\Delta) \quad (4\text{-}6)$$

其中$E(\cdot)$为触发器编码算法。编码时对触发器图像 Δ 通过一定的透明值进行变换。

以类似的方式，傅立叶图像可以重新转换到空间域。对在频域添加触发器后的频域图$F(u,v)'$进行傅里叶逆变换转换为空间域的后门中毒样本，用如下公式进行：

$$F(m,n)^{t}=\frac{1}{WH}\sum_{u=0}^{W-1}\sum_{v=0}^{H-1}F(u,v)'e^{j2\pi(^{um}/_{W}+^{vn}/_{H})} \quad (4\text{-}7)$$

其中$F(u,v)'$是将触发器图像 Δ 添加到干净样本频域中得到初步的后门样本，$u\in\{0,\dots,W-1\}$、$v\in\{0,\dots,H-1\}$。然后，对于D_1中的所有样本，重复上述步骤，应用公式(4-6)中的算法，得到图像隐写触发器设置$D_{trigger}=\{(F(m,n)_i^t,y^t)\}_{i=1}^{|D_1|}$。详细的隐写触发器生成过程在算法 4-1 中进行了总结。

§4.3.3 嵌入后门

使用上述方法生成有毒训练集$D_{trigger}$后，攻击者会利用它将干净子集D_1替换掉，并且按照公式 4-3 更新总的训练数据集D_{train}。当模型开发人员接触到此类来源不可靠的数据集，即使用D_{train}训练模型时后门可以被嵌入，模型训练时采用交叉熵损失的标准模型训练过程，即解决以下优化问题:

$$\underset{\theta}{\operatorname{argmin}}-\frac{1}{N}\sum_{f(m,n)\in D_{train}}\sum_{i=1}^{C}y_i\log\big(p_i(f(m,n),\theta)\big) \quad (4\text{-}8)$$

式中，y_i为$f(m,n)$的对应标签的第i个值，p_i为F_B的$Softmax$的第i个输出，θ为可训练模型权值。优化式可以用 SGD（随机梯度下降）优化器的反向传播来求解。

由于数据集D_{train}中包含了比值为 κ的有毒的数据，因此模型可以学习从触发器

到目标标签的映射，即在训练过程中后门会被嵌入到模型中。在推理预测阶段，攻击者可以通过将触发器注入到良性样本输入中并将其输入到后门感染模型中来激活后门行为。

算法 4-1 ISBA 算法伪代码

算法 4-1：ISBA 触发器生成算法伪代码

输入:干净训练样本集 $\boldsymbol{D}=\{(\boldsymbol{f}(\boldsymbol{m},\boldsymbol{n})_{\boldsymbol{i}},\boldsymbol{y}_{\boldsymbol{i}})\}_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}$，中毒样本注入率 κ ，触发器 Δ。

输出:后门触发器训练样本集（中毒数据集）$\boldsymbol{D}_{\boldsymbol{trigger}}$

1：在干净训练样本集$\boldsymbol{D}$中随机选取比例κ的数据样本放入$\boldsymbol{D}_{\boldsymbol{1}}$

2：**for all** $\boldsymbol{f}(\boldsymbol{m},\boldsymbol{n})_{\boldsymbol{i}} \in \boldsymbol{D}_{\boldsymbol{1}}$ **do**

3：　利用二维离散傅里叶变换将干净样本从空间域$\boldsymbol{f}(\boldsymbol{m},\boldsymbol{n})_{\boldsymbol{i}}$转换到频域$\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})$

4：　将后门触发器 Δ 添加到频域$\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})$中得到 $\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})'$

5：　将 $\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})'$ 利用逆离散傅里叶变换从频域转换到空间域得到 $\boldsymbol{F}(\boldsymbol{m},\boldsymbol{n})^{\boldsymbol{t}}$

6：　将 $\boldsymbol{F}(\boldsymbol{m},\boldsymbol{n})^{\boldsymbol{t}}$加入后门触发器训练样本集$\boldsymbol{D}_{\boldsymbol{trigger}}$中

7：**end for**

8：返回后门中毒样本集 $\boldsymbol{D}_{\boldsymbol{trigger}}$

§4.4 实验结果分析

为了评估本节提出的基于图像隐写触发器的后门攻击方法在攻击效果和攻击鲁棒性等方面的性能，本小节使用两种不同的基准数据集 GTSRB 和 ImageNet、两种不同的神经网络结构 VGG16 和 ResNet18 进行了大量的实验。选取两种最流行的有毒标签后门方法，Gu 等人[16]提出的 BadNets 和 Chen 等人[28]提出的 Blending，作为本节提出方法的参考基准。

§4.4.1 实验参数设置

本章节遵循了第三章中的数据集选取和网络选取，评估了此后门攻击在两个基准数据集上的性能，分别为 ImageNet 和 GTSRB。同样从两个数据集中随机选择一个连续的包含 12 个类别的子集，用于神经网络模型的训练和测试，其中第一个类别被定义为目标类。对于 ImageNet 和 GTSRB，所选的子集分别包含 15,592 张图片和 40,520 张图片。将两个子集按 10:1 的比例分成训练集和测试集，并采用数据增强方法（随机裁剪和旋转）对样本进行处理。这些图像都被调整为 244×244×3。

所有对这两个数据集的攻击都是在 ResNet18 和 VGG16 上进行的，每个数据集分别在两个模型上训练了 200 个轮次，ImageNet 数据集大概需要 23 个小时完成训练、GTSRB 大约需要 13 个小时完成训练。注射速率默认为 $\kappa = 0.09$。在训练阶段使用

SGD 优化器，初始学习速率设置为 0.01。批量大小和最大迭代分别设置为 32 和 200。我们使用的评价指标包括 ASR（攻击成功率）、ATA（攻击后准确度）和P_{BA}。

表 4-1 不同后门方法的性能，其中x/y表示平均指标ATA/ASR，最好的结果用粗体表示

Dataset	Model	$\boldsymbol{BTA}$	Blending[28]	BadNets[16]	ISBA
GTSRB	ResNet18	**93.87**	90.04/99.80	93.05/99.13	93.70/**100**
GTSRB	VGG16	92.31	92.83/99.97	**93.22**/97.39	92.45/**100**
ImageNet	ResNet18	87.30	85.12/99.32	84.43/97.24	**87.32**/**100**
ImageNet	VGG16	86.90	85.34/99.46	84.13/92.05	**89.17**/**99.94**

§4.4.2 结果分析

本节从四个角度讨论了基于图像隐写的后门攻击算法性能，分别为：准确性、攻击有效性、隐秘性、鲁棒性。

（1）准确性：它的目的是测试后门模型遇到干净测试样本时，对干净样本的预测能力是否会因为嵌入后门而受到影响。作为比较，我们使用两种干净训练数据集分别训练上述网络架构 VGG16、ResNet18 以得到四种良性模型。此外，我们还使用 Blending 和 BadNets 方法在以上数据集和网络结构上训练后门模型，精度结果如表 4-1 所示。将我们的 ATA和良性模型的 BTA值相互比较，得到 P_{BA}值，由表 4-2 可以看出，我们的方法只有在 GTSRB 加 ResNet18 组合下该值为负，即此时后门模型对良性测试样本的预测准确率较低于良性模型，但很明显，在遇到干净样本时，在大部分情况下我们的后门攻击对模型预测性能没有负面影响。甚至，在 VGG16 上训练的后门模型对于干净样本的预测准确度有轻微的提高。但是，从表 4-2 中第四和第五列可以看出，Blending 和 BadNets 的预测准确性总体保存得不是很好。例如，在 Blending 后门模型中，使用 GTSRB 训练的 ResNet18 模型的 ATA下降了 3.83%；而在 BadNets 后门模型中，用 ImageNet 训练的 ResNet18 模型的 ATA下降了 2.87%。总之，基于图像隐写的后门攻击触发器 ISBA 实现了高准确性。

表 4-2 不同后门方法在 ImageNet 和 GTSRB 数据集上的$P_{BA}(\%)$，最好的结果用粗体表示

Dataset	Model	***BTA***	Blending[28]	BadNets[16]	ISBA
GTSRB	ResNet18	93.87	-3.83	-0.82	**-0.17**
GTSRB	VGG16	92.31	0.52	**0.91**	0.14
ImageNet	ResNet18	87.30	-2.18	-2.87	**0.02**
ImageNet	VGG16	86.90	-1.56	-2.77	**2.27**

（2）攻击有效性：攻击有效性的目的是量化包含特定触发器的实例激活目标标签的成功率。从表 4-1 可以看出，对于 ISBA 和 Blending 方法，它们的 ASR均能达到或接近 100%。而对于使用 VGG16 在 ImageNet 上训练的 BadNets，ASR约为 92%，且基于 BadNets 的后门攻击方法只有使用 GTSRB 数据集在 ResNet18 上训练得到的后门感染模型 ASR超过了 99%，其余后门感染模型能够达到的攻击成功率均较低。这意味着基于 BadNets 的后门方法其攻击有效性相对于 ISBA 和 Blending 方法来说是较差的，可能是后门触发器在样本上的区域占比导致了这一结果，且可以合理猜测当开发者使用更大的裁剪力度对由 BadNets 生成的包含触发器的训练集进行数据增强操作时，BadNets 相较于 ISBA 和 Blending 方法的 ASR会更低。而本章节提出的 ISBA 方法在四种后门感染模型上的 ASR都比 Blending 更高，这可能是由于本节提出的方法相较于 Blending 方法得到的中毒样本在经过旋转和裁剪等一系列数据增强操作后，触发器变形程度更小。综上所述，本节提出的 ISBA 方法具有高攻击有效性。

（3）隐秘性：隐秘性是用来衡量这些中毒样本引起模型开发者或使用者怀疑的可能性有多大。从直观上看，被嵌入后门触发器的图像样本越自然，中毒样本数量上注入率越小，被污染的数据就越隐秘，模型开发人员就越不可能注意到它们。图 4-1 中显示了通过 BadNets、Blending 以及 ISBA 方法生成的后门样本实例。很明显，由 Blending 和 BadNets 创建的后门样本实例，其触发器模式较干净样本的内容相比相关性较差，有明显人工合成的痕迹。相比之下，使用 ISBA 创建的中毒样本看起来更自然，由于触发器模式的不可见性使得其与宿主样本的内容相关性更强，中毒样本与干净样本相比内容不受影响，中毒样本不易引起模型开发者或使用者怀疑。另一方面，ISBA 可以在相对较低的注入率下实现高准确性和攻击有效性。例如，在 GTSRB 和 ImageNet 上训练的 ISBA 在后门样本注入率为 0.09 时最高能达到 100%的攻击成功率。因此，可以得出 ISBA 满足隐秘性标准的结论。

（4）鲁棒性：后门模型的鲁棒性评估的目的是衡量后门方法是否能够承受后门防御，在经过后门防御后的后门依然能够有效。在本节中，我们主要关注的是 1）微调防御和 2）基于梯度的类激活图防御。

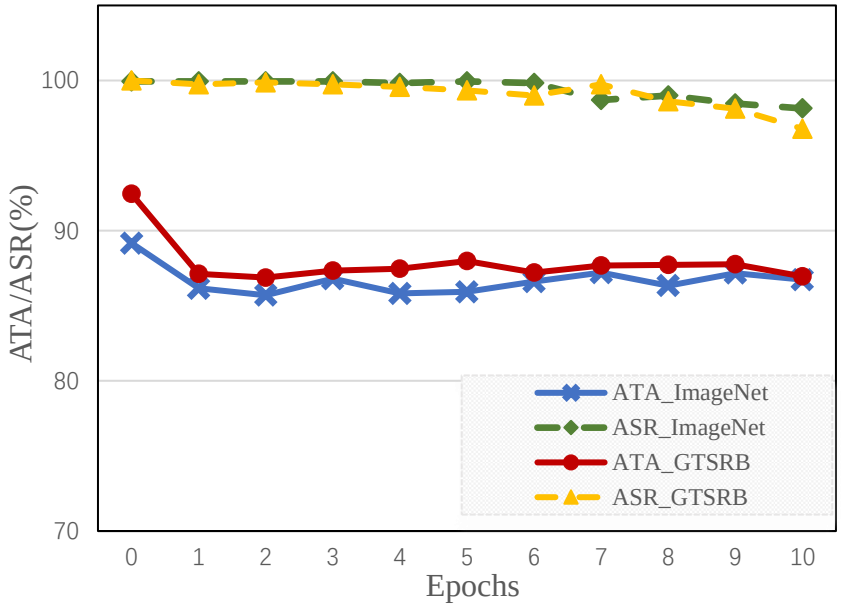


图 4-3 基于图像隐写的后门攻击（ISBA）模型微调

1）微调防御（Fine-tuning）。我们评估了 Blending、BadNets 和 ISBA 在抵御微调防御方面的效果。利用这三种攻击方法，使用 ImageNet 和 GTSRB 数据集在 VGG16 模型上对后门模型进行预训练，然后对它们进行 10 个周期的微调，学习率设置为 0.001，用于微调的数据集是干净的 10%的 ImageNet 和 GTSRB 数据集，微调结果如图 4-3、图 4-4、图 4-5 所示。由图 4-5 可见，BadNets 的 ASR在仅经过 2 个 epoch 的微调后就呈现显著下降趋势，且随着微调周期的不断增大，ASR持续下降。最后，经过 10 个周期的微调，BadNets 的 ASR从下降近 60%。相反，随着微调周期数的增加，Blending 和 ISBA 的 ASR受影响程度小，即使经过 10 个周期的微调，其后门攻击有效性依然在 95%以上。ISBA 方法在 GTSRB 数据集上的的 ASR在微调周期为 10 时下降为了约 96%，而此设置下的 Blending 方法 ASR 最终维持在 97%左右，此时 Blending 方法较优于 ISBA 方法。而 ISBA 方法在 ImageNet 数据集上的 ASR在微调周期为 10 时下降为了 98.14%，在此设置下的 Blending 方法 ASR最终下降为 97.01%，此时 ISBA 方法反而更优于 Blending。总的来说，ISBA 可以与 Blending 方法相媲美，而且两者在抵御微调防御方面都优于 BadNets。这可能是由于 BadNets 的触发器过于简单，触发器模式只关注图像的一小部分，预测干净输入的神经元很少与预测触发器的神经元重叠，由此，BadNets 更容易受到微调防御的影响。

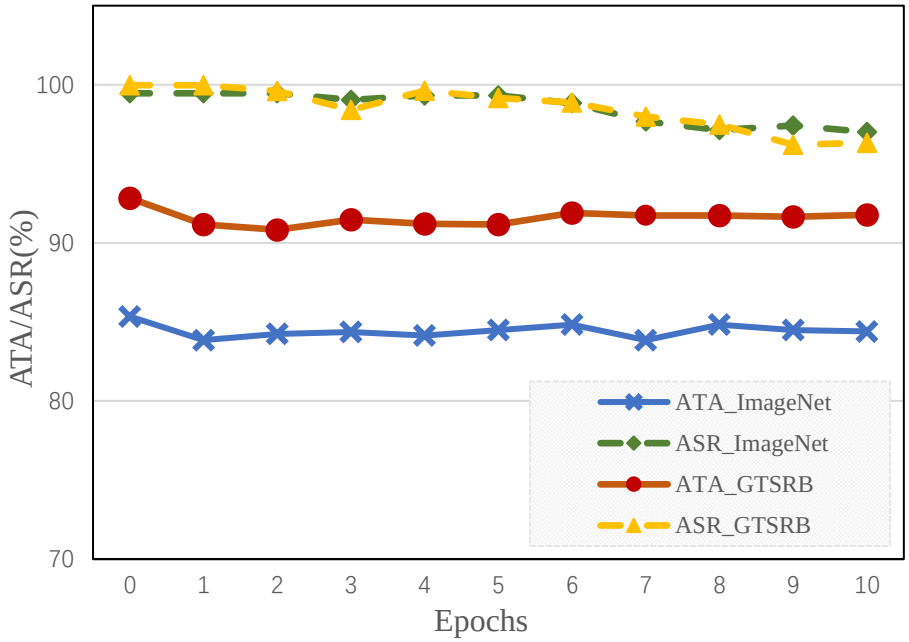


图 4-4 基于 Blending 的后门攻击模型微调

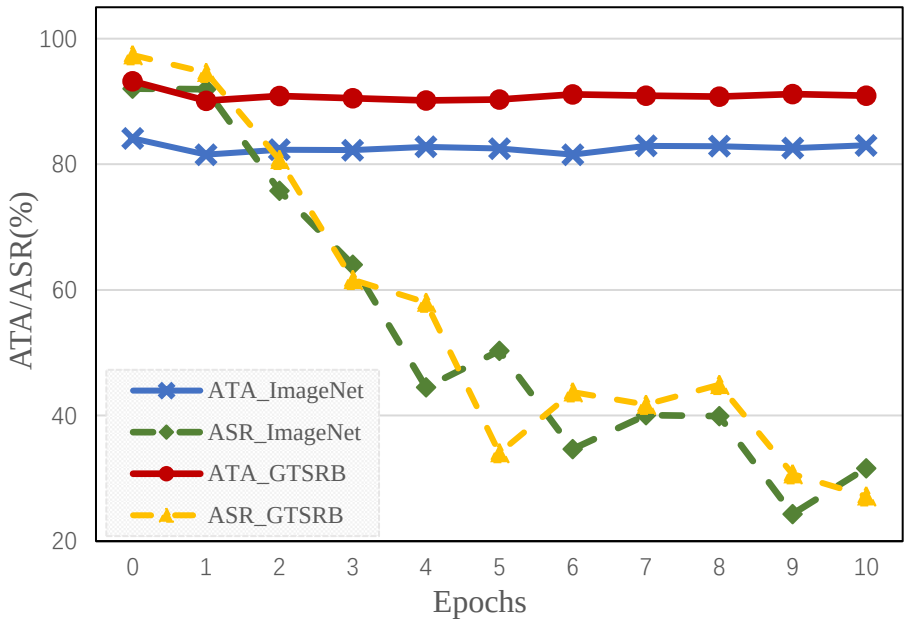


图 4-5 基于 BadNet 的后门攻击模型微调

2）基于梯度的类激活图防御（Grad-CAM）。如第二章所述，Grad-CAM 防御方法的有效性高度依赖于恶意触发器显著区域的定位精度。为了评估我们的方法对这种防御的抵抗性，我们在 VGG16 上对给定的数据集 ImageNet 和 GTSRB，通过 Grad-CAM 生成中毒样本的类激活图，如图 4-6、图 4-7 所示。从以上两个图中第一行可以看出，由 BadNets 获得的后门中毒样本的类激活图激活部分集中在特定的显著区域，即图像的右下角，这正是该方法后门触发器嵌入的位置。这种过于集中的显著区域很可能被识别为恶意触发器区域。对于 Blending 生成的类激活图，如两图第二行所示，它们的突出显著的区域不像 BadNets 方法那样聚焦在图像的小部分固定区域上。而由 ISBA

生成的中毒样本的显著激活区域，如图第三行所示，高亮区域相较于 BadNets 同样更为分散。这主要是因为基于 ISBA 生成的有毒图像包含的触发器和 Blending 一样，相较于 BadNets 而言，在样本实例上的区域更大更为分散，而由 BadNets 生成的有毒图像具有位置和区域较为固定的触发器模式。因此，与 BadNets 相比，本方法生成的触发器更难区分，且不逊于 Blending。综上所述，该攻击对以基于梯度的类激活图防御有一定的抵抗能力。

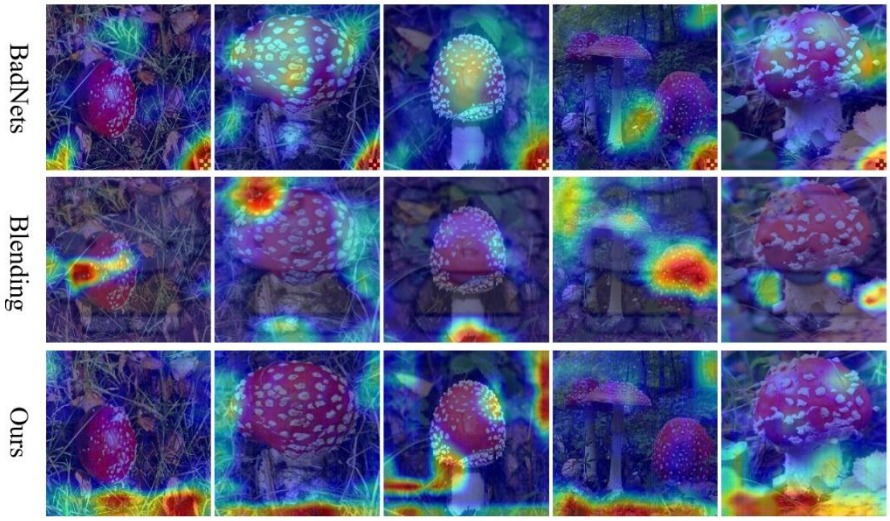


图 4-6 Grad-CAM 实例_ImageNet

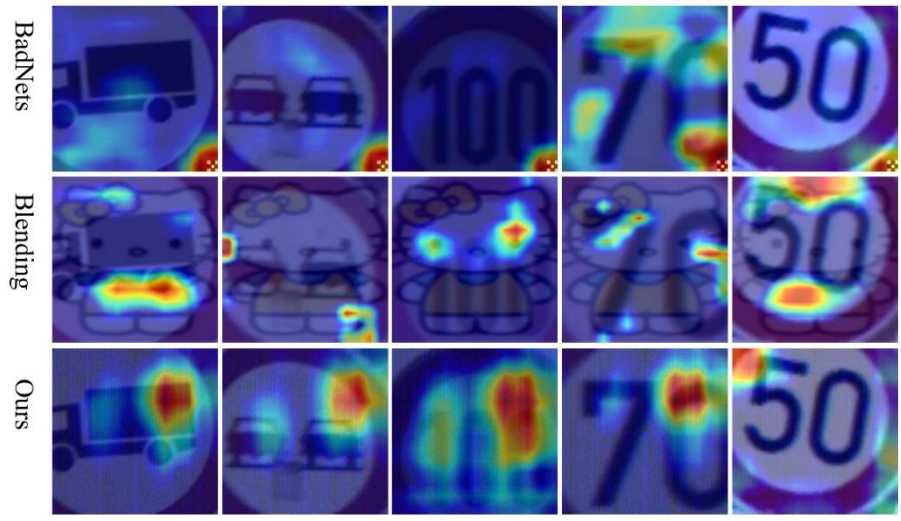


图 4-7 Grad-CAM 实例_GTSRB

§4.5 本章小结

在本章节介绍了我们提出的基于图像隐写触发器的后门攻击方法。说明了由于大多数现有的后门攻击方法其触发器模式都是受限的或触发器内容与其宿主样本内容不相关联的模式，这不仅使后门样本由于其不自然的外观很容易被模型开发人员怀疑，而且还允许当前的后门防御轻松地缓解后门攻击行为。基于此，本工作可以在保证后

门行为的同时将后门自然地嵌入模型，与现有的后门触发器相比，该攻击方法更加隐秘，可以规避数据过滤。而且除了后门感染模型在推理预测阶段，模型内隐藏的后门能够被触发器恶意触发外，攻击者还可以通过解码的手段对训练数据集发起验证。此外，在第三章节提出的基于雨滴触发器的后门攻击方法不能兼容所有类型的数据集，但在本节提出的方法具有对于不同类型数据集的兼容性更强的优势。这使得该方法对现有的后门防御更具抵抗力。本章节通过大量的实验，验证了基于图像隐写的后门攻击方法在攻击 VGG16 和 ResNet18 这两种不同神经网络模型时在准确性、攻击有效性、隐秘性和鲁棒性方面的效果。

第五章 总结与展望

本文研究工作主要围绕基于自然触发器的后门攻击方法展开，所使用的研究方法理论主要包括了基于梯度下降 SGD 的模型优化算法、卷积神经网络、离散傅里叶变换等，主要的研究成果为基于雨滴触发器的后门攻击方法和基于图像隐写的不可见触发器后门攻击方法。

§5.1 工作总结

本文主要针对后门攻击展开研究，后门攻击指攻击者意图在 DNN 中注入隐藏的后门，使被攻击模型在良性样本上表现良好，并且可通过由攻击者设定好的后门触发器影响模型预测。后门攻击已成为深度神经网络模型的严重威胁，针对其的相关研究可以在深度学习安全方面起到重要的作用。但是目前，最流行有效的后门触发器往往是在固定位置使用同一个触发器处理不同的干净数据，或嵌入触发器时不考虑宿主数据的内容导致与宿主样本内容相关性差，此外，中毒样本可能是通过简单地将触发器与良性宿主样本直接叠加而生成。在实际的应用中，这些触发器并不能满足在神经网络中隐秘地嵌入后门。为此，本文针对以上问题，提出了两种基于自然后门触发器的后门攻击算法。主要研究内容如下：

（1）利用自然行为，提出了一种用于图像分类任务的后门攻击方法:基于雨滴触发器的后门攻击算法（RDBA）。首先通过随机噪声与 α值保证雨滴触发器的均匀随机分布以及控制触发器密度，再通过构建好的对角矩阵与旋转矩阵进行仿射变换得到一个对角核，再对该核进行高斯模糊且使用该核对初步生成的噪声图进行滤波操作，使得触发器完成由从最初的随机噪声图到有宽度、长度、运动模糊的雨滴图的转变。然后，我们将雨滴触发器与一小部分干净的训练样本合并，生成看起来自然的有毒数据。最后，通过多分类网络模型训练得到后门模型，并通过在 ImageNet 和 GTSRB 数据集上的大量实验证明了该方法在使用当前流行的防御机制攻击模型时的鲁棒性、攻击有效性、隐秘性等。

（2）提出了一种基于图像隐写的不可见触发器后门攻击方法。首先，通过二维快速傅里叶变换将由攻击者选取的含有目标类别信息的图片作为触发器，将其在频域嵌入少量干净样本中，通过将在频域添加好触发器的中毒样本进行逆傅里叶变换得到最终的中毒样本。再将其注入剩余干净样本中得到攻击样本训练集。其次，将后门攻击样本与干净样本混合共同送入多分类神经网络训练，以得到后门模型。最后，在 ImageNet 和 GTSRB 数据集上用大量实验证明了该方法在面对当前流行的防御机制

攻击模型时的攻击有效性、鲁棒性等。

§5.2 未来展望

（1）关于第三章基于雨滴触发器的后门攻击方法研究，由于现实中很难收集成对的自然雨滴图像，大多数需要使用雨滴的工作使用人工生成的数据集，比如部分去雨工作会使用包括 Rain100L 和 Rain100H 在内的人工数据集。所以，第三章中合成的后门触发器与现实中的雨天图像并不全然相同。由于对于一个预算有限的模型开发人员来说，手工检查每个训练示例是很困难的，所以该方法依然能成功欺骗开发者将后门模型部署成功且大量实验也证明了该方法的攻击有效性、隐秘性以及在抵御当前流行的防御机制时的鲁棒性。但在未来的工作中仍然可以考虑通过更复杂的基于深度学习的解决方案来解决如何让人工合成的雨滴图像更加真实这个问题，关键是如何自动生成与真实雨滴无法区分的合成雨滴。一些先进的技术，如生成对抗网络或风格转移网络，可以用于这一目的。其中攻击者和开发人员的能力可以作出改进（这意味着攻击者需要有强大的计算能力和足够的标记数据，它们都需要包含真实雨滴的数据集来训练网络）。本文认为对它们的探索是高级攻击/防御值得单独研究，这项研究将在未来深入。

（2）本文在第三、四章节提出的两种方法主要针对机器视觉图像处理上的后门攻击进行研究，然而深度学习随着时间的推移越来越被广泛应用于物联网的多个领域，各个领域也均应对后门攻击这类安全问题引起高度重视,例如语音领域等。语音领域从初期发展到现在衍生出了多类工作，应用于我们生活的各个方面，其安全重要性同计算机视觉领域一样需要引起更高的重视，因此未来将会考虑更多领域相对应的后门攻击研究。

（3）本文所讨论的两种后门攻击均为基于有毒标签的后门攻击，此类后门攻击对攻击者需要掌握的神经网络结构信息要求相比于目前基于干净标签的后门攻击而言较低，因而更容易在实际生活中应用且攻击成功率往往较高。但是基于干净标签的后门攻击因其不用改变中毒样本的实际真实标签特性，后门隐秘性的上限更高，故未来将会考虑对基于干净标签的后门攻击工作进行探讨。

（4）目前本文探讨了两种不同的后门攻击模式，然而针对各类后门攻击的新防御也是层出不穷，目前的后门攻击方法终将会因为更多具有针对性的后门防御方法的出现而逐渐失效。可以说后门攻击与后门防御两者之间相互促进彼此的发展，因此对后门攻击的研究是一个需要长期不断努力课题。

参考文献

[1] Caesar H, Bankiti V, Lang A, et al. NuScenes: a multimodal dataset for autonomous driving[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11618–11628.

[2] Chen C, Seff A, Kornhauser A, et al. DeepDriving: learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 2722-2730.

[3] Urmson C, Baker C, Dolan J, et al. Autonomous driving in traffic: boss and the urban challenge[C]//Proceedings of the Conference on Innovative Applications Artificial Intelligence, 2008, 30(2): 17-28.

[4] Szegedy C, Vincent V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 2818–2826.

[5] Dai T, Cai J, Zhang Y, et al. Second order attention network for single image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11065–11074.

[6] Dalbelo Bǎsić B, Buono M P. An analysis of early use of deep learning terms in natural language processing[C]//Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology, 2020: 1125–1129.

[7] Devlin J, Chang M.-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171–4186.

[8] Rao S, Daume H. Learning to ask good questions: ranking clarification questions using neural expected Value of perfect information[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2018: 2737–2746.

[9] Goodfellow I, Shlens J, and Szegedy C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations, 2015.

[10] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the International Conference on Learning Representations, 2018.

[11] Nguyen A, Tran A. Wanet–Imperceptible warping based backdoor attack[C]//Proceedings of the International Conference on Learning Representations, 2021.

[12] Mustafa A, Khan S, Hayat M, et al. Image super-resolution as a defense against adversarial

attacks[J]. IEEE Transactions on Image Processing, 2020, 29: 1711–1724.

[13] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses[C]//Proceedings of the International Conference on Learning Representations, 2018.

[14] Li B, Vorobeychik Y. Scalable optimization of randomized operational decisions in adversarial classification settings[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics, 2015: 599–607.

[15] Dai H, Li H, Tian T, et al. Adversarial attack on graph structured data[C]//Proceedings of the International Conference on Machine Learning, 2018: 1115–1124.

[16] Gu T, Liu K, Dolan-Gavitt B, et al. BadNets: evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.

[17] Lin S, Zhang Y, Hsu C, et al. The architectural implications of autonomous driving: constraints and acceleration[C]//Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2018: 751–766.

[18] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016: 1528–1540.

[19] Wright J, Ganesh A, Zhou Z, et al. Demo: robust face recognition via sparse representation[C]//Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008: 942-943.

[20] Liu C. The development trend of evaluating face-recognition technology[C]//Proceedings of the 2014 International Conference on Mechatronics and Control, 2014: 1540-1544.

[21] Heigold, Moreno I, Bengio S, et al. End-to-end text-dependent speaker verification[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2016: 5115–5119.

[22] Snyder D, Garcia-Romero D, Sell G, et al. Speaker recognition for multi-speaker conversations using x-vectors[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2019: 5796–5800.

[23] Snyder D and Garcia-Romero D, Sell G, et al. X-Vectors: robust DNN embeddings for speaker recognition[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2018: 5329–5333.

[24] Kemker R, McClure M, Abitino A, et al. Measuring catastrophic forgetting in neural networks[C]//Proceedings of the Association for the Advance of Artificial Intelligence, 2018.

[25] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Proceedings of the Recent Advances in Intrusion Detection, 2018: 273–294.

[26] Selvaraju R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via

gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 618–626.

[27] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: generalized gradient based visual explanations for deep convolutional networks[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018: 839–847.

[28] Chen X, Chang L, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv:1712.05526, 2017.

[29] Ali S, Huang W, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS), 2018.

[30] Zhu C, Huang W, Li G, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proceedings of the 2019 International Conference on Machine Learning, 2019: 7614–7623.

[31] Pan S.J, Qiang Y. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[32] Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks[C]//Proceedings of Network and Distributed System Security Symposium, 2017.

[33] 刘全,翟建伟,章宗长,钟珊,周倩,章鹏,徐进.深度强化学习综述[J]. 计算机学报, 2018, 41(01): 1-27.

[34] Yang Z, Iyer N, Reimann J, et al. Design of intentional backdoors in sequential models[J]. arXiv preprint arXiv:1902.09972, 2019.

[35] Kiourti P, Wardega K, Jha S, et al. TrojDRL: trojan attacks on deep reinforcement learning agents[J]. arXiv preprint arXiv:1903.06638, 2019.

[36] Mcmahan H.B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017: 1273-1282.

[37] Konen J, McMahan H.B, Yu F.X, et al. Federated learning: strategies for improving communication efficiency[J]. arXiv preprint arXiv:1610.05492, 2016.

[38] Geyer R.C, T Klein, M Nabi. Differentially private federated learning: a client level perspective[C]//Proceedings of the 2017 Neural Information Processing Systems, 2017.

[39] Rehak D.R, P Dodds, L Lannom. A model and infrastructure for federated learning content repositories. [C]//Proceedings of the International World Wide Web Conference, 2005.

[40] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning?[C]//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 2020: 2938-2947.

[41] Sun Z, Kairouz P, Suresh A.T, et al. Can You Really Backdoor Federated Learning?[J]. arXiv preprint arXiv: 1911.07963, 2019.

[42] Chou E, Tramèr F, Pellegrino G. SentiNet: detecting localized universal attacks against deep learning systems[C]//Proceedings of the 2020 IEEE Security and Privacy Workshops, 2020: 48–54.

[43] Huang X, Alzantot M, Srivastava M. NeuronInspect: detecting backdoors in neural networks via output explanations[J]. arXiv preprint arXiv: 1911.07399, 2019.

[44] 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): 1004-9037.

[45] 周飞燕, 金林鹏, 董军, 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.

[46] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014.

[47] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations, no. 1556, ICLR, 2015.

[48] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.

[49] Jia D, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[50] Stallkamp J, Schlipsing M, Jan S, et al. The german traffic sign recognition benchmark: a multi-class classification competition[C]//Proceedings of the International Joint Conference on Neural Networks, 2011: 1453–1460.

[51] Wang B, Yao Y, Shan S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of the IEEE Symposium on Security and Privacy, 2019: 707–723.

[52] Ruder S. An overview of gradient descent optimization algorithms[J]. arXiv preprint arXiv:1609.04747, 2016.

[53] Namias V. The Fractional Order Fourier Transform and its Application to Quantum Mechanics[J]. Ima Journal of Applied Mathematics, 1980, 25: 241-265.

致谢

天波易谢，寸暑难留。转眼硕士研究生三年的学习即将结束，回首过往，收获颇多。在这里，要向耐心给予教育和指导的所有老师以及给予关心陪伴的所有小伙伴们表示深深的感谢!

首先要感谢我的导师赵峰老师以及 Leo Yu Zhang 老师和蓝如师老师在学业上给予我的诸多帮助，论文写作过程中，从选题立意、撰写修改直至最终定稿都是在几位导师的悉心指导下完成的。三年的学习生活中，三位导师给予了我很多做人和学习方面的教诲，导师严谨的学风、深厚的学术功底让人从内心深处感到钦佩，这些都将激励我在今后的学习、工作中严格要求自己，不断努力。这里，要向导师们表示由衷的感谢和真诚的敬意。

感谢人工智能交叉研究院的所有同门们三年来的关怀与协助。首先感谢钟绮师姐与王小琴师姐对于我论文上的帮助，在你们身上我学到了严谨认真的科研态度。以及感谢师兄师姐，郑金云，臧美美，赵文婷，孙龙，王文颢，王如月，张莉，胡锡普，朱奕杰，关善文，朱煜，陈曦，周李，黄珈瑜，王中帅等给予我的照顾。然后我要感谢实验室小伙伴陈晨、卞小曼、黄锴、秦子钦以及同门张华楠、陶训芳、李猷兴、黄惠文、叶天鸽、孙波、张家伟、唐丽玲、唐文博、许广在日常学习以及生活中给予我陪伴、温暖与快乐。同时，感谢师弟师妹，谭钰、周靖淞、刘智轩、焦志勇、戴六连、陈颖贤、魏陈浩、李航为我的生活增添新的活力与快乐。最后感谢我的室友秦子钦、李淼、李静怡在宿舍生活中给予我的照顾。希望今后我们仍能像现在一样，互相帮助，共同学习和成长。

衷心地感谢在百忙之中参加论文评审和答辩的各位专家和教授。

最后，感谢我的父母及男友给予我支持与爱。

作者在攻读硕士期间的主要研究成果

论文

[1] Feng Zhao, **Li Zhou**, Qi Zhong, Rushi Lan, and Leo Yu Zhang. Natural Backdoor Attacks on Deep Neural Networks via Raindrops[J]. Security and Communication Networks. (SCI 收录，DOI: 10.1155/2022/4593002) (对应于本文第三章)