29.All current backdoor attacks on deep learning (DL) models fall underthe category of a vertical class backdoor (VCB)—class-dependent.In VCB attacks, any sample from a class activates the implantedbackdoor when the secret trigger is present, regardless of whetherit is a sub-type source-class-agnostic backdoor or a source-class-specific backdoor. For example, a trigger of sunglasses can mis-lead a facial recognition model into administrator prediction whenany people (source-class-agnostic) or a specific group of people(source-class-specific) wear sunglasses. Existing defense strategiesoverwhelmingly focus on countering VCB attacks, especially thosethat are source-class-agnostic. This narrow focus neglects the po-tential threat of other simpler yet general backdoor types, leadingto false security implications.It is, therefore, crucial to discoverand elucidate unknown backdoor types, particularly those thatcan be easily implemented, as a mandatory step before developingcountermeasures.This study introduces a new, simple, and general type of back-door attack coined as the horizontal class backdoor (HCB) thattrivially breaches the class dependence characteristic of the VCB,bringing a fresh perspective to the community. HCB is now ac-tivated when the trigger is presented together with an innocu-ous feature, regardless of class.For example, the facial recognitionmodel misclassifies a person who wears sunglasses with a smilinginnocuous feature into the targeted person, such as an adminis-trator, regardless of which person. Smiling is innocuous becauseit is irrelevant to the main task of facial recognition. The key isthat these innocuous features (such as rain, fog, or snow in au-tonomous driving or facial expressions like smiling or sadness infacial recognition) are horizontally shared among classes but are only exhibited by partial samples per class. Extensive experimentson attacking performance across various tasks, including MNIST,facial recognition, traffic sign recognition, object detection, andmedical diagnosis,confirm the high efficiency and effectivenessof the HCB. We rigorously evaluated the evasiveness of the HCBagainst a series of eleven representative countermeasures, includ-ing Fine-Pruning (RAID 18’), STRIP (ACSAC 19’), Neural Cleanse(Oakland 19’), ABS (CCS 19’), Februus (ACSAC 20’), NAD (ICLR21’), MNTD (Oakland 21’), SCAn (USENIX SEC 21’), MOTH (Oak-land 22’), Beatrix (NDSS 23’), and MM-BD (Oakland 24’). None ofthese countermeasures prove robustness, even when employing asimplistic trigger, such as a small and static white-square patch.

机器学习即服务（MLaaS）是一种满足市场需求的新兴产品。然而，用户在使用MLaaS时需要将数据上传到远程服务器，这引发了隐私担忧。自“被遗忘权”生效以来，云上产品已广泛支持数据删除功能，以便从远程数据集和机器学习模型中移除用户的私人数据。近期，提出了许多机器学习数据删除方法来消除被遗忘数据的影响。不幸的是，我们发现机器学习数据删除会使云上模型极易受到后门攻击。在本文中，我们报告了一种针对启用数据删除功能的模型的新威胁，并实现了一种基于影响驱动的伪装的后门攻击（Unlearning Activated Backdoor Attack with Influence-driven camouflage，简称UBA-Inf）。不同于传统的后门攻击，UBA-Inf通过机器学习数据删除激活伪装的后门，提供了一种新的后门方法，以提高攻击的有效性和隐蔽性。所提出的方法可以使用现成的后门生成算法来实现。此外，UBA-Inf是一种“按需”攻击，通过数据删除请求提供对后门激活的细粒度控制，克服了后门消失和暴露的问题。通过对UBA-Inf进行广泛评估，我们得出结论，UBA-Inf是一种强大的后门方法，提高了隐蔽性、鲁棒性和持久性。

28.Machine-Learning-as-a-Service (MLaaS) is an emergingproduct to meet the market demand. However, end users are re-quired to upload data to the remote server when using MLaaS,raising privacy concerns. Since the right to be forgotten cameinto effect, data unlearning has been widely supported in on-cloud products for removing users’ private data from remotedatasets and machine learning models. Plenty of machineunlearning methods have been proposed recently to erasethe influence of forgotten data. Unfortunately, we find thatmachine unlearning makes the on-cloud model highly vul-nerable to backdoor attacks. In this paper, we report a newthreat against models with unlearning enabled and implementan Unlearning Activated Backdoor Attack with Influence-driven camouflage (UBA-Inf). Unlike conventional backdoorattacks, UBA-Inf provides a new backdoor approach for effec-tiveness and stealthiness by activating the camouflaged back-door through machine unlearning. The proposed approachcan be implemented using off-the-shelf backdoor generatingalgorithms. Moreover, UBA-Inf is an “on-demand” attack,offering fine-grained control of backdoor activation throughunlearning requests, overcoming backdoor vanishing and ex-posure problems. By extensively evaluating UBA-Inf, weconclude that UBA-Inf is a powerful backdoor approach thatimproves stealthiness, robustness, and persistence.

基于显著性的表示可视化（SRV）（如Grad-CAM）因其简单性和高效性，成为最经典且广泛采用的可解释人工智能（XAI）方法之一。它可以通过定位对预测贡献最大的显著区域来解释深度神经网络。然而，由于样本的真实显著区域缺乏，自动测量和评估SRV方法的性能变得困难。在本文中，我们重新审视了基于后门的SRV评估方法，这是目前唯一可行的缓解上述问题的方法。我们首先揭示了其由于现有后门水印的触发泛化能力而存在的实现局限性和不可靠性。基于这些发现，我们提出了一种泛化受限的后门水印（GLBW），并基于此设计了一种更可靠的XAI评估方法。具体来说，我们将带水印DNN的训练表述为一个最小-最大问题，其中我们通过内部最大化找到“最坏”的潜在触发器（具有最高的攻击有效性和与真实触发器的最大差异），并通过每次迭代中的外部最小化来最小化其影响以及良性样本和中毒样本上的损失。特别是，我们设计了一种自适应优化方法，以在每次内部最大化中找到所需的潜在触发器。在基准数据集上进行了大量实验，验证了我们的泛化受限水印的有效性。

27.Saliency-based representation visualization (SRV) (e.g.,Grad-CAM) is one ofthe most classical and widely adopted explainable artificial intelligence (XAI)methods for its simplicity and efficiency. It can be used to interpret deep neu-ral networks by locating saliency areas contributing the most to their predictions.However, it is difficult to automatically measure and evaluate the performance ofSRV methods due to the lack of ground-truth salience areas of samples. In thispaper, we revisit the backdoor-based SRV evaluation, which is currently the onlyfeasible method to alleviate the previous problem. We first reveal its implementa-tion limitations and unreliable nature due to the trigger generalization of existingbackdoor watermarks. Given these findings, we propose a generalization-limitedbackdoor watermark (GLBW), based on which we design a more faithful XAIevaluation. Specifically, we formulate the training of watermarked DNNs as amin-max problem, where we find the ‘worst’ potential trigger (with the highestattack effectiveness and differences from the ground-truth trigger) via inner max-imization and minimize its effects and the loss over benign and poisoned samplesvia outer minimization in each iteration. In particular, we design an adaptive op-timization method to find desired potential triggers in each inner maximization.Extensive experiments on benchmark datasets are conducted, verifying the ef-fectiveness of our generalization-limited watermark.

在后门攻击中，攻击者将恶意构造的后门样本插入训练集，使得生成的模型易受操控。防御此类攻击的方法包括将插入的样本视为训练集中的异常值，并使用鲁棒统计学技术来检测和移除它们。在本文中，我们提出了一种不同的方法来应对后门攻击问题。具体来说，我们证明了在没有训练数据分布的结构信息的情况下，后门攻击与数据中的自然特征无法区分——因此，在一般意义上“检测”它们是不可能的。然后，基于这一观察，我们重新审视了现有的后门攻击防御措施，并描述了它们所做出的（通常是隐含的）假设，以及它们所依赖的条件。最后，我们从另一个角度探讨了后门攻击：即假设这些攻击对应于训练数据中最强的特征。在这个假设（我们将其形式化）下，我们开发了一种新的基本方法来检测后门攻击。我们的基本方法自然地产生了一种带有理论保证且在实践中有效的检测算法。

26.In a backdoor attack,an adversary inserts ma-liciously constructed backdoor examples into atraining set to make the resulting model vulner-able to manipulation. Defending against suchattacks involves viewing inserted examples as out-liers in the training set and using techniques fromrobust statistics to detect and remove them.In this work, we present a different approach tothe backdoor attack problem. Specifically, weshow that without structural information aboutthe training data distribution, backdoor attacksare indistinguishable from naturally-occuring fea-tures in the data—and thus impossible to “detect”in a general sense. Then, guided by this observa-tion, we revisit existing defenses against backdoorattacks and characterize the (often latent) assump-tions they make, and on which they depend. Fi-nally, we explore an alternative perspective onbackdoor attacks: one that assumes these attackscorrespond to the strongest feature in the train-ing data. Under this assumption (which we makeformal) we develop a new primitive for detectingbackdoor attacks. Our primitive naturally givesrise to a detection algorithm that comes with the-oretical guarantees, and is effective in practice.

人工智能安全领域的一大目标是安全可靠地生产和部署用于实际应用的深度学习模型。为此，近年来广泛探索了生产阶段（或训练阶段）深度神经网络（DNN）中的数据投毒基后门攻击及相应的防御措施。具有讽刺意味的是，部署阶段的后门攻击，这类攻击常发生在非专业用户的设备上，因此在现实场景中可以说威胁更大，却并未引起社区的太多关注。我们将这种警惕性的不平衡归因于现有部署阶段后门攻击算法的实用性较弱以及现实世界中攻击演示的不足。为了填补这一空白，本文研究了部署阶段对DNN的后门攻击所带来的现实威胁。我们的研究基于一种常用的部署阶段攻击范式——对抗性权重攻击，其中攻击者选择性地修改模型权重以在部署的DNN中嵌入后门。为了实现现实的实用性，我们提出了第一个用于后门注入的灰盒且物理可实现的权重攻击算法，即子网替换攻击（SRA），该算法仅需要受害者模型的架构信息，并且可以在现实世界中支持物理触发器。我们进行了大量的实验模拟和系统级的现实世界攻击演示。我们的结果不仅表明了所提出攻击算法的有效性和实用性，还揭示了一种新型计算机病毒的实际风险，这种病毒可能会广泛传播，并在用户设备中秘密地向DNN模型注入后门。通过我们的研究，我们呼吁更多关注DNN在部署阶段的脆弱性。

25.One major goal of the AI security community is to securelyand reliably produce and deploy deep learning models forreal-world applications. To this end, data poisoning basedbackdoor attacks on deep neural networks (DNNs) in theproduction stage (or training stage) and correspondingdefenses are extensively explored in recent years. Ironi-cally, backdoor attacks in the deployment stage, which canoften happen in unprofessional users’ devices and are thusarguably far more threatening in real-world scenarios,draw much less attention of the community. We attributethis imbalance of vigilance to the weak practicality ofexisting deployment-stage backdoor attack algorithms andthe insufﬁciency of real-world attack demonstrations.Toﬁll the blank, in this work, we study the realistic threat ofdeployment-stage backdoor attacks on DNNs. We baseour study on a commonly used deployment-stage attackparadigm — adversarial weight attack, where adversariesselectively modify model weights to embed backdoorinto deployed DNNs. To approach realistic practicality,we propose the ﬁrst gray-box and physically realizableweights attack algorithm for backdoor injection, namelysubnet replacement attack (SRA),which only requiresarchitecture information of the victim model and cansupport physical triggers in the real world. Extensiveexperimental simulations and system-level real-worldattack demonstrations are conducted. Our results not onlysuggest the effectiveness and practicality of the proposedattack algorithm, but also reveal the practical risk ofanovel type of computer virus that may widely spreadand stealthily inject backdoor into DNN models in userdevices.By our study, we call for more attention to thevulnerability of DNNs in the deployment stage.

含有后门的深度学习模型在被触发时会表现出恶意行为，但在其他情况下看似正常。这种风险常因模型外包而增加，对模型的安全使用构成了挑战。尽管已存在一些对策，但它们在防御适应性攻击方面的效果尚未得到充分检验，这可能导致对安全性的误判。本研究是首次深入探讨在外包模型中检测后门的难度，尤其是当攻击者调整其策略时，即使其能力受到显著限制。攻击者可以通过简单地违反威胁模型（例如，使用未被检测覆盖的高级后门类型或触发器设计）来规避检测，这相对容易实现。然而，本研究强调，即使在其定义的威胁模型和有限的对抗能力下（例如，使用易于检测的触发器同时保持高攻击成功率），各种主要的检测防御也可以同时被简单的适应性策略规避。更具体地说，本研究介绍了一种新方法，该方法以共生的方式采用触发器特异性增强和训练调节。这种方法使我们能够同时规避多种后门检测防御，包括Neural Cleanse（Oakland 19'）、ABS（CCS 19'）和MNTD（Oakland 21'）。这些是2022年NeurIPS木马检测挑战赛“规避型木马”赛道所选用的检测工具。即使在严格条件下（如高攻击成功率（>97%）和仅限使用最简单的触发器（小白方块））与这些防御措施结合使用时，我们的简单方法也获得了NeurIPS木马检测挑战赛的二等奖。值得注意的是，我们的适应性攻击首次成功规避了其他最新的先进防御措施，包括FeatureRE（NeurIPS 22'）和Beatrix（NDSS 23'）。本研究表明，现有的模型外包后门防御措施在适应性攻击面前仍然脆弱，因此，应尽可能避免使用第三方模型。

24.Deep learning models with backdoors act mali-ciously when triggered but seem normal otherwise. This risk,often increased by model outsourcing, challenges their secure use.Although countermeasures exist, their defense against adaptiveattacks is under-examined, possibly leading to security misjudg-ments. This study is the first intricate examination illustrating thedifficulty of detecting backdoors in outsourced models, especiallywhen attackers adjust their strategies, even if their capabilitiesare significantly limited. It is relatively straightforward forattackers to circumvent detection by trivially violating its threatmodel (e.g., using advanced backdoor types or trigger designsnot covered by the detection). However, this research highlightsthat various leading detection defenses can simultaneously beevaded using simple adaptive strategies, even under their definedthreat models and with limited adversary capabilities (e.g.,using easily detectable triggers while maintaining a high attacksuccess rate). To be more specific, this study introduces a novelmethodology that employs trigger specificity enhancement andtraining regulation in a symbiotic manner. This approach allowsus to evade multiple backdoor detection defenses simultaneously,including Neural Cleanse (Oakland 19’), ABS (CCS 19’), andMNTD (Oakland 21’). These were the detection tools selected forthe Evasive Trojans Track of the 2022 NeurIPS Trojan DetectionChallenge. Even when applied in conjunction with these defensesunder stringent conditions, such as a high attack success rate(> 97%) and the restricted use of the simplest trigger (small whitesquare), our straightforward method garnered the second prize inNeurIPS Trojan Detection Challenge. Notably, for the first time,our adaptive attack successfully evaded other recent state-of-the-art defenses, including FeatureRE (NeurIPS 22’) and Beatrix(NDSS 23’). This study suggests that existing model outsourcingbackdoor defenses remain vulnerable to adaptive attacks, and thus, the use of third-party models should be avoided wheneverpossible.

扩散模型（DMs）是最先进的生成模型，它们通过迭代添加噪声和去噪的过程学习一种可逆的损坏过程。它们是许多生成式人工智能应用（如文本到图像的条件生成）的核心。然而，最近的研究表明，基本的无条件扩散模型（例如，DDPM [16]和DDIM [52]）易受后门注入攻击，这是一种由模型输入中恶意嵌入的模式触发的输出操纵攻击。本文提出了一种统一的后门攻击框架（VillanDiffusion），以扩展当前针对扩散模型的后门分析范围。我们的框架涵盖了主流的无条件和条件扩散模型（基于去噪和基于得分的模型）以及各种无需训练的采样器，以进行全面评估。实验表明，我们的统一框架有助于对不同配置的扩散模型进行后门分析，并为针对扩散模型的基于标题的后门攻击提供了新的见解。

23.Diffusion Models (DMs) are state-of-the-art generative models that learn a re-versible corruption process from iterative noise addition and denoising. They arethe backbone of many generative AI applications, such as text-to-image conditionalgeneration. However, recent studies have shown that basic unconditional DMs(e.g., DDPM [16]and DDIM [52]) are vulnerable to backdoor injection, a type ofoutput manipulation attack triggered by a maliciously embedded pattern at modelinput. This paper presents a unified backdoor attack framework (VillanDiffusion)to expand the current scope of backdoor analysis for DMs. Our framework coversmainstream unconditional and conditional DMs (denoising-based and score-based)and various training-free samplers for holistic evaluations. Experiments show thatour unified framework facilitates the backdoor analysis of different DM configura-tions and provides new insights into caption-based backdoor attacks on DMs.

随着机器学习技术的迅速传播，共享和采用公共机器学习模型变得非常流行。这为攻击者提供了许多新机会。在本文中，我们提出了一种针对神经网络的木马攻击。由于模型对人类来说并不直观，因此这种攻击具有隐蔽性。部署被植入木马的模型可能导致各种严重后果，包括危及人类生命（如在自动驾驶等应用中）。我们首先逆向神经网络以生成一个通用木马触发器，然后使用逆向工程的训练数据重新训练模型，以向模型中注入恶意行为。这些恶意行为仅当输入被加盖木马触发器时才会被激活。在我们的攻击中，我们无需篡改原始训练过程，该过程通常需要数周至数月的时间。相反，我们的攻击只需几分钟到几小时的时间。此外，我们也不需要用于训练模型的数据集。在实践中，由于隐私或版权问题，数据集通常不会共享。我们使用五个不同的应用来演示我们攻击的强大威力，并对可能影响攻击的因素进行了深入分析。结果表明，我们的攻击高度有效且高效。被植入木马的行为可以成功触发（几乎达到100%的可能性），同时不会影响模型对正常输入的测试准确率，甚至在公共数据集上的准确率更高。此外，攻击一个复杂的神经网络模型只需少量时间。最后，我们还讨论了针对此类攻击的可能防御措施。

22.With the fast spread of machine learning tech-niques, sharing and adopting public machine learning models be-come very popular. This gives attackers many new opportunities.In this paper, we propose a trojaning attack on neural networks.As the models are not intuitive for human to understand, theattack features stealthiness. Deploying trojaned models can causevarious severe consequences including endangering human lives(in applications like autonomous driving). We ﬁrst inverse theneural network to generate a general trojan trigger,and thenretrain the model with reversed engineered training data to injectmalicious behaviors to the model. The malicious behaviors areonly activated by inputs stamped with the trojan trigger. In ourattack, we do not need to tamper with the original trainingprocess, which usually takes weeks to months. Instead, it takesminutes to hours to apply our attack. Also, we do not require thedatasets that are used to train the model. In practice, the datasetsare usually not shared due to privacy or copyright concerns.We use ﬁve different applications to demonstrate the power ofour attack, and perform a deep analysis on the possible factorsthat affect the attack. The results show that our attack is highlyeffective and efﬁcient. The trojaned behaviors can be successfullytriggered (with nearly 100% possibility) without affecting its testaccuracy for normal input and even with better accuracy onpublic dataset. Also, it only takes a small amount of time toattack a complex neuron network model. In the end, we alsodiscuss possible defense against such attacks.

随着深度神经网络（DNN）在众多应用中的普及，这些网络的安全性变得至关重要。预训练的DNN可能包含通过中毒训练注入的后门。这些被植入木马的模型在提供常规输入时表现良好，但当输入被加盖一个称为木马触发器的独特模式时，会误分类为目标输出标签。近期，针对基于DNN的人工智能应用，提出了各种后门检测和缓解系统。然而，其中许多系统仅限于需要特定补丁触发器的木马攻击。在本文中，我们介绍了一种更灵活且隐蔽的木马攻击——复合攻击，它利用由多个标签的现有良性特征组成的木马触发器来躲避后门扫描器。我们表明，具有复合后门的神经网络在良性数据上的准确率可与原始版本相媲美，而当输入中存在复合触发器时，则会发生误分类。我们在7个不同任务上的实验表明，这种攻击构成了严重威胁。我们使用两种最先进的后门扫描器对我们的攻击进行了评估。结果显示，两种扫描器都无法检测到注入的任何后门。我们还详细研究了扫描器为何无效。最后，我们讨论了本次攻击的本质，并提出了可能的防御措施。

21.With the prevalent use of Deep Neural Networks (DNNs) in manyapplications, security of these networks is of importance. Pre-trained DNNs may contain backdoors that are injected throughpoisoned training. These trojaned models perform well when regu-lar inputs are provided, but misclassify to a target output label whenthe input is stamped with a unique pattern called trojan trigger. Re-cently various backdoor detection and mitigation systems for DNNbased AI applications have been proposed. However, many of themare limited to trojan attacks that require a specific patch trigger.In this paper, we introduce composite attack,a more flexible andstealthy trojan attack that eludes backdoor scanners using trojantriggers composed from existing benign features of multiple labels.We show that a neural network with a composed backdoor canachieve accuracy comparable to its original version on benign dataand misclassifies when the composite trigger is present in the input.Our experiments on 7 different tasks show that this attack posesasevere threat. We evaluate our attack with two state-of-the-artbackdoor scanners. The results show none of the injected backdoorscan be detected by either scanner. We also study in details why thescanners are not effective. In the end, we discuss the essence of ourattack and propose possible defense.

深度跨模态哈希因其卓越的效率和存储优势，促进了多模态检索领域的发展，但其对后门攻击的脆弱性却鲜有研究。值得注意的是，当前的深度跨模态哈希方法不可避免地需要大量训练数据，这导致含有难以察觉触发器的恶意样本可以轻易地伪装成训练数据，从而在受害模型中埋入后门。然而，现有的后门攻击主要集中在单模态视觉领域，而多模态差距和哈希量化削弱了它们的攻击性能。为应对上述挑战，本文针对深度跨模态哈希检索展开了一种隐形的黑盒后门攻击。据我们所知，这是该研究领域的首次尝试。具体而言，我们开发了一种灵活的触发器生成器，用于生成攻击者指定的触发器，该生成器学习非中毒模态的样本语义，以弥合跨模态攻击的差距。然后，我们设计了一个输入感知注入网络，该网络以样本特定的隐蔽形式将生成的触发器嵌入良性样本中，并实现触发器与中毒样本之间的跨模态语义交互。由于受害模型的知识不可知，我们使得任何跨模态哈希仿制品都能促进黑盒后门攻击，并减轻哈希量化对攻击的削弱作用。此外，我们还提出了一种混淆扰动和掩码策略，以诱导高性能的受害模型关注中毒样本中的难以察觉的触发器。在基准数据集上的大量实验表明，我们的方法对深度跨模态哈希检索具有最先进的攻击性能。此外，我们还研究了可迁移攻击、少样本中毒、多模态中毒、可察觉性以及潜在防御对后门攻击的影响。

20.Deep cross-modal hashing has promoted the field of multi-modal retrieval due to its excellent efficiency andstorage, but its vulnerability to backdoor attacks is rarely studied. Notably, current deep cross-modal hash-ing methods inevitably require large-scale training data, resulting in poisoned samples with imperceptibletriggers that can easily be camouflaged into the training data to bury backdoors in the victim model. Nev-ertheless, existing backdoor attacks focus on the uni-modal vision domain, while the multi-modal gap andhash quantization weaken their attack performance. In addressing the aforementioned challenges, we un-dertake an invisible black-box backdoor attack against deep cross-modal hashing retrieval in this article. Tothe best of our knowledge, this is the first attempt in this research field. Specifically, we develop a flexibletrigger generator to generate the attacker’s specified triggers, which learns the sample semantics of the non-poisoned modality to bridge the cross-modal attack gap. Then, we devise an input-aware injection network,which embeds the generated triggers into benign samples in the form of sample-specific stealth and realizescross-modal semantic interaction between triggers and poisoned samples. Owing to the knowledge-agnosticof victim models, we enable any cross-modal hashing knockoff to facilitate the black-box backdoor attack andalleviate the attack weakening of hash quantization. Moreover, we propose a confusing perturbation and maskstrategy to induce the high-performance victim models to focus on imperceptible triggers in poisoned sam-ples. Extensive experiments on benchmark datasets demonstrate that our method has a state-of-the-art attackperformance against deep cross-modal hashing retrieval. Besides, we investigate the influences of transfer-able attacks, few-shot poisoning, multi-modal poisoning, perceptibility, and potential defenses on backdoorattacks.

深度神经网络（DNN）因其在各种应用中的广泛使用，其安全性日益受到关注。近期，已部署的DNN被证明易受木马攻击，这种攻击通过比特翻转操纵模型参数，注入隐藏行为，并通过特定触发模式激活。然而，所有现有的木马攻击都采用了显眼的基于补丁的触发器（例如，方形图案），使得它们对人类肉眼可见，也容易被机器检测到。在本文中，我们提出了一种新型攻击，即难以察觉的木马攻击（HPT）。HPT通过利用加性噪声和逐像素光流场，分别调整原始图像的像素值和位置，从而制作出难以察觉的木马图像。为了实现更优的攻击性能，我们提出联合优化比特翻转、加性噪声和光流场。由于DNN的权重比特是二进制的，这个问题非常难以解决。我们通过等价替换处理二进制约束，并提供了一种有效的优化算法。在CIFAR-10、SVHN和ImageNet数据集上的大量实验表明，所提出的HPT能够生成难以察觉的木马图像，同时与最先进的方法相比，实现了相当或更好的攻击性能。

19.The security of deep neural networks (DNNs) has attractedincreasing attention due to their widespread use in various applications.Recently, the deployed DNNs have been demonstrated to be vulnerableto Trojan attacks, which manipulate model parameters with bit ﬂips toinject a hidden behavior and activate it by a speciﬁc trigger pattern.However, all existing Trojan attacks adopt noticeable patch-based trig-gers (e.g., a square pattern), making them perceptible to humans andeasy to be spotted by machines. In this paper, we present a novel attack,namely hardly perceptible Trojan attack (HPT). HPT crafts hardly per-ceptible Trojan images by utilizing the additive noise and per-pixel ﬂowﬁeld to tweak the pixel values and positions of the original images, respec-tively. To achieve superior attack performance, we propose to jointlyoptimize bit ﬂips, additive noise, and ﬂow ﬁeld. Since the weight bitsof the DNNs are binary, this problem is very hard to be solved. Wehandle the binary constraint with equivalent replacement and providean eﬀective optimization algorithm. Extensive experiments on CIFAR-10, SVHN, and ImageNet datasets show that the proposed HPT cangenerate hardly perceptible Trojan images, while achieving comparableor better attack performance compared to the state-of-the-art methods.

在后门攻击中，攻击者向训练集中注入被破坏的样本。攻击者的目标是使得最终训练好的模型在测试输入中添加预定义触发器时，预测出攻击者期望的目标标签。这些攻击的核心在于攻击成功率与注入的被破坏训练样本数量之间的权衡。我们将这种攻击构建为一个新型的双层优化问题：构造能够最大化训练模型攻击成功率的强效投毒样本。我们使用神经正切核来近似被攻击模型的训练动态，并自动学习强效的投毒样本。我们在CIFAR-10和ImageNet的子类上，使用WideResNet-34和ConvNeXt架构，针对周期性触发和补丁触发攻击进行了实验，结果表明，与基线相比，NTBA设计的投毒样本在注入数量减少十倍的情况下，仍能实现90%的攻击成功率。我们使用核线性回归的分析对NTBA设计的攻击进行了解释。我们还进一步展示了过参数化深度神经网络中的一个漏洞，这一漏洞由神经正切核的形状所揭示。

18.In a backdoor attack, an attacker injects corrupted examples into the training set. The goal of theattacker is to cause the ﬁnal trained model to predict the attacker’s desired target label when a predeﬁnedtrigger is added to test inputs. Central to these attacks is the trade-oﬀ between the success rate of theattack and the number of corrupted training examples injected. We pose this attack as a novel bileveloptimization problem: construct strong poison examples that maximize the attack success rate of thetrained model. We use neural tangent kernels to approximate the training dynamics of the model beingattacked and automatically learn strong poison examples. We experiment on subclasses of CIFAR-10and ImageNet with WideResNet-34 and ConvNeXt architectures on periodic and patch trigger attacksand show that NTBA-designed poisoned examples achieve, for example, an attack success rate of 90%with ten times smaller number of poison examples injected compared to the baseline. We provided aninterpretation of the NTBA-designed attacks using the analysis of kernel linear regression. We furtherdemonstrate a vulnerability in overparametrized deep neural networks, which is revealed by the shape ofthe neural tangent kernel.

基于数据投毒的后门攻击旨在通过操纵训练数据集，在不控制目标模型训练过程的情况下，向模型中插入后门。现有的攻击方法主要集中在设计触发器或触发器与良性样本之间的融合策略。然而，这些方法通常随机选择样本进行投毒，忽视了每个投毒样本在后门注入方面的重要性差异。最近的一种选择策略通过记录遗忘事件来过滤固定大小的投毒样本池，但它未能从全局角度考虑池外剩余样本。此外，计算遗忘事件需要大量的额外计算资源。因此，如何高效且有效地从整个数据集中选择投毒样本，是后门攻击中亟待解决的问题。为了解决这一问题，我们首先在常规的后门训练损失中引入了投毒掩码。我们假设，使用难以学习的投毒样本训练的后门模型对易于学习的样本具有更强的后门效应，这可以通过阻碍正常训练过程（即相对于掩码最大化损失）来实现。为了将其进一步融入正常训练过程，我们提出了一种可学习的投毒样本选择策略，通过最小化-最大化优化，同时学习掩码和模型参数。具体来说，外层循环旨在通过最小化基于所选样本的损失来实现后门攻击目标，而内层循环则通过最大化损失来选择阻碍这一目标的难以学习的投毒样本。经过几轮对抗性训练后，我们最终选出了具有高贡献度的有效投毒样本。在基准数据集上的大量实验证明了我们的方法在提升后门攻击性能方面的有效性和效率。

17.Data-poisoning based backdoor attacks aim to insert backdoor into models bymanipulating training datasets without controlling the training process of the targetmodel. Existing attack methods mainly focus on designing triggers or fusion strate-gies between triggers and benign samples. However, they often randomly selectsamples to be poisoned, disregarding the varying importance of each poisoningsample in terms of backdoor injection. A recent selection strategy filters a fixed-sizepoisoning sample pool by recording forgetting events, but it fails to consider the re-maining samples outside the pool from a global perspective. Moreover, computingforgetting events requires significant additional computing resources. Therefore,how to efficiently and effectively select poisoning samples from the entire datasetis an urgent problem in backdoor attacks. To address it, firstly, we introduce a poi-soning mask into the regular backdoor training loss. We suppose that a backdooredmodel training with hard poisoning samples has a more backdoor effect on easyones, which can be implemented by hindering the normal training process (i.e.,maximizing loss w.r.t. mask). To further integrate it with normal training process,we then propose a learnable poisoning sample selection strategy to learn the masktogether with the model parameters through a min-max optimization. Specifically,the outer loop aims to achieve the backdoor attack goal by minimizing the lossbased on the selected samples, while the inner loop selects hard poisoning samplesthat impede this goal by maximizing the loss. After several rounds of adversar-ial training, we finally select effective poisoning samples with high contribution.Extensive experiments on benchmark datasets demonstrate the effectiveness andefficiency of our approach in boosting backdoor attack performance.

深度神经网络（DNN）容易受到后门攻击，即嵌入恶意功能，使攻击者能够触发错误的分类。传统的后门攻击使用强触发器特征，这些特征很容易被受害模型学习。尽管这种攻击对输入变化具有鲁棒性，但这种鲁棒性却增加了无意触发激活的可能性。这会给现有的防御手段留下痕迹，这些防御手段通过反向工程和样本叠加等方法，可以找到原始触发器的近似替代品，这些替代品能够在不与原始触发器完全相同的情况下激活后门。在本文中，我们提出并研究了一种后门攻击的新特性，即后门排他性，它衡量的是后门触发器在输入变化存在时保持有效性的能力。基于后门排他性的概念，我们提出了后门排他性提升（BELT）技术，这是一种新颖的技术，通过抑制后门与模糊触发器之间的关联来增强后门排他性，从而规避防御。在三个流行的后门基准上的广泛评估验证了我们的方法，该方法显著增强了四种传统后门攻击的隐蔽性，在经过后门排他性提升后，这些攻击几乎在不牺牲攻击成功率和正常效用的情况下，能够规避七种最先进的后门对策。例如，经BELT增强的最早的后门攻击之一BadNet，能够规避大多数最先进的防御措施，包括ABS和MOTH，否则这些防御措施会识别出被植入后门的模型。

16.Deep neural networks (DNNs) are susceptible tobackdoor attacks, where malicious functionality is embedded toallow attackers to trigger incorrect classifications. Old-schoolbackdoor attacks use strong trigger features that can easilybe learned by victim models. Despite robustness against inputvariation, the robustness however increases the likelihood ofunintentional trigger activations. This leaves traces to existingdefenses, which find approximate replacements for the originaltriggers that can activate the backdoor without being identicalto the original trigger via, e.g., reverse engineering and sampleoverlay.In this paper, we propose and investigate a new char-acteristic of backdoor attacks, namely, backdoor exclusivity,which measures the ability of backdoor triggers to remaineffective in the presence of input variation. Building upon theconcept of backdoor exclusivity, we propose Backdoor Exclu-sivity LifTing (BELT), a novel technique which suppressesthe association between the backdoor and fuzzy triggers toenhance backdoor exclusivity for defense evasion. Extensiveevaluation on three popular backdoor benchmarks validate,our approach substantially enhances the stealthiness of fourold-school backdoor attacks, which, after backdoor exclusiv-ity lifting, is able to evade seven state-of-the-art backdoorcountermeasures, at almost no cost of the attack success rateand normal utility. For example, one of the earliest backdoorattacks BadNet, enhanced by BELT, evades most of the state-of-the-art defenses including ABS and MOTH which wouldotherwise recognize the backdoored model.

深度神经网络（DNN）易受后门攻击，这种攻击会污染训练集，从而改变模型对带有特定触发器的样本的预测。虽然现有研究主要集中在单模态场景，但现代人工智能系统通常采用多种模态来提高模型性能，这使得多模态后门攻击更加实用，但由于模态间的固有交互、多个攻击面、模态贡献不均衡等因素，其结构也更为复杂。这些因素对多模态学习中的后门攻击的有效性有显著影响，但尚未得到充分研究。为弥补这一空白，我们提出了首个针对多模态学习的数据高效且计算高效的后门攻击方法。我们的解决方案包含两项创新。首先，我们提出了一种新颖的基于后门梯度的评分（BAGS），它可以在训练早期阶段准确量化每个数据样本对后门学习的贡献。因此，它可以为攻击者节省大量的时间和计算资源。其次，我们引入了一种具有两种攻击模式的搜索策略，以高效地确定最佳的投毒模态和数据样本。我们的方法带来了以下研究成果。首先，我们在最先进的多模态任务、模型、数据集和设置上对提出的解决方案进行了全面评估，以验证其有效性、效率和可迁移性。例如，我们只需投毒0.005%的训练样本，即可成功攻击视觉问答任务，成功率超过96%。对于音视频语音识别任务，我们投毒0.05%的样本，即可实现超过93%的成功率。其次，我们在实验过程中发现了几个有趣的发现：（1）投毒所有模态并不总是比单独投毒一个模态效果更好，有时甚至会使攻击效果更差；（2）多模态学习后门攻击中存在着模态竞争和互补性并存的现象；（3）多模态学习中的主导模态可能并不主导后门攻击。我们希望这项工作能激发未来对提高多模态学习安全性的进一步研究。

15.Deep Neural Networks (DNNs) are vulnerable tobackdoor attacks, which poison the training set to alter themodel prediction over samples with a specific trigger. Whileexisting efforts mainly focus on unimodal scenarios, modernAI systems usually employ multiple modalities to improvethe model performance, making multimodal backdoor attacksmore practical but structurally more complex due to inherentmodality interactions, multiple attack surfaces, unbalancedmodality contributions, etc. These factors affect the effective-ness of backdooring multimodal learning significantly but havenot been fully investigated yet.To bridge this gap, we present the first data and computa-tion efficient backdoor attacks towards multimodal learning.Our solution consists of two innovations. First, we proposeanovel backdoor gradient-based score (BAGS), which canaccurately quantify the contribution of each data sample to thebackdoor learning at a very early training stage. Therefore,it can greatly save time and computational resources for theattacker. Second, we introduce a searching strategy with twoattack modes to efficiently determine the optimal poisoningmodalities and data samples.Our methodology leads to the following research outcomes.First, we comprehensively evaluate the proposed solution overstate-of-the-art multimodal tasks, models, datasets and settings,to verify its effectiveness, efficiency and transferability. Forinstance, we only need to poison 0.005% of training samples toattack the Visual Question Answering task with the success rateof >96%. For the Audio Video Speech Recognition task, wepoison 0.05% of samples to achieve the success rate of >93%.Second, we disclose several interesting findings during ourexperiments: (1) poisoning all modalities is not always betterthan individual ones, sometimes even making the attack worse;(2) modality competition and complementarity coexist in mul-timodal learning backdoor attacks; (3) A dominant modality inmultimodal learning may not dominate the backdoor attacks.We hope this work will spur future research in improvingthe security of multimodal learning.

联邦学习（FL）是一种事实上的分布式机器学习范式，它在个人设备上本地训练数据集，但易受后门模型投毒攻击。攻击者通过攻破或冒充这些设备，可以上传精心制作的恶意模型更新，从而在攻击者指定的触发条件下，操纵全局模型表现出后门行为。然而，现有的后门攻击需要了解受害者FL系统的更多信息，这超出了实际的黑盒设置。此外，它们通常专注于优化单一目标，而随着现代FL系统倾向于采用从不同角度检测后门模型的深度防御措施，这种单一目标的攻击变得无效。基于这些考虑，本文提出了3DFed，一个自适应、可扩展且多层的框架，用于在黑盒设置中发起隐蔽的FL后门攻击。3DFed具备三个规避模块来伪装后门模型：带约束损失的后门训练、噪声掩码和诱饵模型。通过在后门模型中植入指示器，3DFed可以从全局模型中获得前一个训练轮次的攻击反馈，并动态调整这些后门规避模块的超参数。通过大量的实验结果，我们表明，当所有组件协同工作时，3DFed可以规避所有最先进的FL后门防御措施的检测，包括Deepsight、Foolsgold、FLAME、FL-Detector和RFLBAT。由于3DFed是一个可扩展的框架，未来还可以纳入新的规避模块。

14.Federated Learning (FL), the de-facto distributedmachine learning paradigm that locally trains datasets at indi-vidual devices, is vulnerable to backdoor model poisoning attacks.By compromising or impersonating those devices, an attacker canupload crafted malicious model updates to manipulate the globalmodel with backdoor behavior upon attacker-speciﬁed triggers.However, existing backdoor attacks require more informationon the victim FL system beyond a practical black-box setting.Furthermore, they are often specialized to optimize for a singleobjective, which becomes ineffective as modern FL systemstend to adopt in-depth defense that detects backdoor modelsfrom different perspectives. Motivated by these concerns, in thispaper, we propose 3DFed, an adaptive, extensible, and multi-layered framework to launch covert FL backdoor attacks inablack-box setting. 3DFed sports three evasion modules thatcamouﬂage backdoor models: backdoor training with constrainedloss, noise mask, and decoy model. By implanting indicatorsinto a backdoor model, 3DFed can obtain the attack feedbackin the previous epoch from the global model and dynamicallyadjust the hyper-parameters of these backdoor evasion modules.Through extensive experimental results, we show that when allits components work together, 3DFed can evade the detection ofall state-of-the-art FL backdoor defenses, including Deepsight,Foolsgold, FLAME, FL-Detector, and RFLBAT. New evasionmodules can also be incorporated in 3DFed in the future asit is an extensible framework.

针对深度神经网络（DNN）模型的后门攻击已被广泛研究。针对不同的领域和范式，如图像、点云、自然语言处理、迁移学习等，已经提出了各种攻击技术。将后门嵌入DNN模型最广泛使用的方法是投毒训练数据。他们通常从良性训练集中随机选择样本进行投毒，而没有考虑每个样本对后门有效性的不同贡献，这使得攻击不是最优的。最近的一项工作[40]提出使用遗忘得分来衡量每个投毒样本的重要性，然后过滤掉冗余数据以进行有效的后门训练。然而，这种方法是凭经验设计的，没有理论证明。而且它非常耗时，因为需要经历几个训练阶段来选择数据。为了解决这些限制，我们提出了一种基于置信度的新型评分方法，该方法可以根据距离后验有效地衡量每个投毒样本的贡献。我们还进一步引入了一种贪婪搜索算法，以更迅速地找到最适合注入后门的样本。在二维图像和三维点云分类任务上的实验评估表明，我们的方法可以达到与基于遗忘得分的搜索方法相当的性能，甚至超越它，同时只需要标准训练过程中几个额外的训练周期的计算量。

13.Backdoor attacks against deep neural network (DNN)models have been widely studied. Various attack techniqueshave been proposed for different domains and paradigms,e.g., image, point cloud, natural language processing,transfer learning, etc. The most widely-used way to em-bed a backdoor into a DNN model is to poison the trainingdata. They usually randomly select samples from the benigntraining set for poisoning, without considering the distinctcontribution of each sample to the backdoor effectiveness,making the attack less optimal.Arecent work [40]proposed to use the forgettingscore to measure the importance of each poisoned sam-ple and then filter out redundant data for effective back-door training. However, this method is empirically de-signed without theoretical proofing. It is also verytime-consuming as it needs to go through several train-ing stages for data selection. To address such lim-itations, we propose a novel confidence-based scoringmethodology, which can efficiently measure the contri-bution of each poisoning sample based on the distanceposteriors. We further introduce a greedy search algo-rithm to find the most informative samples for backdoorinjection more promptly. Experimental evaluations onboth 2D image and 3D point cloud classification tasksshow that our approach can achieve comparable perfor-mance or even surpass the forgetting score-based search-ing method while requiring only several extra epochs’ com-putation of a standard training process.

鉴于深度神经网络在各个领域的有效性，神经网络的安全性受到了广泛关注。后门攻击通过投毒部分训练集来诱导模型的恶意行为，至今仍是一个具有挑战性的问题。近期，许多研究提出了不同的嵌入后门方法来提高后门攻击的隐蔽性。然而，降低投毒样本的比例是提高隐蔽性最直接的方法之一。最近的一项研究（过滤与更新策略，FUS）揭示，投毒样本的选择也至关重要，因为不同的样本对网络最终决策边界的贡献不同。具体来说，他们利用训练阶段每个样本的遗忘事件来识别哪些样本将对网络的预测贡献更大。然而，他们搜索方法的训练阶段计算成本高昂且速度缓慢。为了克服这一问题，本文提出了一种基于训练样本高频能量（HFE）的有效样本选择策略，并采用了全局筛选和更新策略，该策略不仅可以实现更高的后门攻击成功率，而且与FUS相比，搜索时间减少了4320倍（12小时对比10秒）。在CIFAR-10、CIFAR-100和ImageNet-10上的大量实验结果表明，我们提出的方法更简单、更快、更高效。

12.Given the effectiveness of deep neural networksin various fields, the security of neural networks has receivedgreat attention. The backdoor attack, which induces maliciousbehaviors of models by poisoning part of the training set,still remains a challenging problem. Many recent efforts haveproposed different ways of embedding backdoors to improve thestealthiness of backdoor attacks. Yet, lowering the percentageof poisoned samples is one of the most direct ways to increasestealthiness. A recent study (Filtering-and-Updating strategy,FUS) has revealed that the sample selection for poisoning isalso crucial, as different samples contribute differently to thefinal decision boundary of the network. Concretely, they utilizeeach sample’s forgetting events during the training stage toidentify which samples will contribute more to the network’sprediction. The training phase of their search method, however,is computationally expensive and slow. To overcome this, in thispaper, we propose an efficient sample selection strategy basedon the high-frequency energy (HFE) of training samples withaglobal screening and updating strategy, which can not onlyachieve a higher backdoor-attack success rate but also reduce thesearching time by a factor of 4320 compared to FUS (12 hoursvs 10 seconds). The extensive experiment results on CIFAR-10,CIFAR-100, and ImageNet-10 have shown that our proposedmethod is much simpler, faster, and more efficient.

关于后门投毒攻击的大量文献研究了使用“数字触发模式”的后门攻击和防御方法。相比之下，“物理后门”使用物理对象作为触发器，最近才被发现，并且在性质上与针对数字触发后门的大多数防御方法有很大不同，足以抵抗这些防御。对物理后门的研究受限于获取包含与目标错误分类共存的物理对象真实图像的大型数据集。构建这些数据集既费时又费力。本研究旨在解决物理后门攻击研究中的可访问性挑战。我们假设，在如ImageNet等流行数据集中，可能已经存在自然共存的物理对象。一旦识别出这些对象，通过对这些数据进行仔细的重新标记，就可以将它们转变为物理后门攻击的训练样本。我们提出了一种方法，可在现有数据集中可扩展地识别这些潜在的触发器子集，以及它们可以投毒的具体类别。我们将这些自然存在的触发器-类别子集称为自然后门数据集。我们的技术成功地在广泛可用的数据集中识别出了自然后门，并生成了在行为上与手动策划数据集上训练的模型等效的模型。

11..Extensive literature on backdoor poison attacks has studied attacks and defensesfor backdoors using “digital trigger patterns.” In contrast, “physical backdoors”use physical objects as triggers, have only recently been identiﬁed, and are quali-tatively different enough to resist most defenses targeting digital trigger backdoors.Research on physical backdoors is limited by access to large datasets containingreal images of physical objects co-located with misclassiﬁcation targets.Buildingthese datasets is time- and labor-intensive.This work seeks to address the challenge of accessibility for research on physicalbackdoor attacks. We hypothesize that there may be naturally occurring phys-ically co-located objects already present in popular datasets such as ImageNet.Once identiﬁed, a careful relabeling of these data can transform them into trainingsamples for physical backdoor attacks. We propose a method to scalably iden-tify these subsets of potential triggers in existing datasets, along with the speciﬁcclasses they can poison. We call these naturally occurring trigger-class subsets nat-ural backdoor datasets.Our techniques successfully identify natural backdoors inwidely-available datasets, and produce models behaviorally equivalent to thosetrained on manually curated datasets.

深度神经网络（DNN）已渗透到众多不同的应用领域，使其成为恶意攻击的诱人目标。DNN尤其容易受到数据投毒攻击的影响。这种攻击可以通过在不改变训练样本真实标签的情况下对其进行投毒，从而变得更具毒性且更难检测。尽管这一要求很实用，但“干净标签”的限制在同时优化攻击的隐蔽性、成功率和投毒模型的实用性方面产生了强烈的冲突和限制。试图规避这些陷阱往往会导致高注入率、嵌入的后门无效、触发器不自然、可迁移性低和/或鲁棒性差。在本文中，我们通过融合不同的数据增强技术来克服这些限制，以用于后门触发器。根据增强方法对感知损失和目标类激活的增强特征显著性的容忍度，通过插值干净样本及其增强版本，迭代调整增强方法的空间强度。我们提出的攻击在不同的网络模型和数据集上进行了全面评估。与最先进的干净标签后门攻击相比，该攻击具有更低的注入率、更隐蔽的投毒样本、更高的攻击成功率和更强的后门缓解抵抗力，同时保持了较高的良性准确率。在对投毒模型进行权重剪枝和量化后，在Intel神经计算棒2边缘AI设备上的实施也展示了相似的攻击成功率。

10.Deep neural networks (DNNs) have permeated intomany diverse application domains, making them attractive tar-gets of malicious attacks. DNNs are particularly susceptible todata poisoning attacks. Such attacks can be made more venomousand harder to detect by poisoning the training samples withoutchanging their ground-truth labels. Despite its pragmatism, theclean-label requirement imposes a stiff restriction and strongconflict in simultaneous optimization of attack stealth, successrate, and utility of the poisoned model. Attempts to circumventthe pitfalls often lead to a high injection rate, ineffective embed-ded backdoors, unnatural triggers, low transferability, and/orpoor robustness. In this paper, we overcome these constraints byamalgamating different data augmentation techniques for thebackdoor trigger. The spatial intensities of the augmentationmethods are iteratively adjusted by interpolating the cleansample and its augmented version according to their toleranceto perceptual loss and augmented feature saliency to target classactivation. Our proposed attack is comprehensively evaluated ondifferent network models and datasets. Compared with state-of-the-art clean-label backdoor attacks, it has lower injectionrate, stealthier poisoned samples, higher attack success rate,and greater backdoor mitigation resistance while preservinghigh benign accuracy. Similar attack success rates are alsodemonstrated on the Intel Neural Compute Stick 2 edge AI deviceimplementation of the poisoned model after weight-pruning andquantization

垂直联邦学习（VFL）促进了多方在模型训练上的合作，每一方都拥有分布式数据集中部分特征的数据。尽管后门攻击已被发现是联邦学习（FL）安全的主要威胁之一，但关于VFL中的后门攻击的研究仍处于初级阶段。现有的VFL后门攻击方法依赖于使用标签推断等方法预测样本的伪标签，这需要大量额外的信息，而这些信息在实际FL场景中并不易获得。为了评估VFL对后门攻击的实际脆弱性，我们提出了一种针对VFL的目标高效清洁后门（TECB）攻击。TECB方法包括两个阶段：i) 清洁后门投毒（CBP）和目标梯度对齐（TGA）。在CBP阶段，攻击者训练一个后门触发器并在VFL训练过程中投毒模型。投毒后的模型在TGA阶段进一步微调，以增强其在复杂多分类任务中的有效性。与现有方法相比，所提出的TECB在仅掌握目标类样本非常有限信息的情况下，实现了高度有效的后门攻击，这在典型的VFL设置中更为实用。实验结果验证了TECB的卓越性能，在仅知晓0.1%目标标签的情况下，在三个广泛使用的数据集（CIFAR10、CIFAR100和CINIC-10）上实现了超过97%的攻击成功率（ASR），优于最先进的攻击方法。这项研究揭示了VFL中潜在的后门风险，有助于在金融、医疗等领域开发安全的VFL应用。

9.Vertical Federated Learning (VFL) facilitates col-laboration on model training among multiple parties, eachowning partitioned features of the distributed dataset. Althoughbackdoor attacks have been found as one of the main threats toFL security, research on backdoor attacks in VFL is still in theinfant stage. Existing methods for VFL backdoor attacks relyon predicting sample pseudo-labels using approaches such aslabel inference, which require substantial additional informationnot readily available in practical FL scenarios. To evaluate thepractical vulnerability of VFL to backdoor attacks, we presentatarget-efﬁcient clean backdoor (TECB) attack for VFL. TheTECB approach consists of two phases – i) Clean BackdoorPoisoning (CBP) and Target Gradient Alignment (TGA). In theCBP phase, the adversary trains a backdoor trigger and poisonsthe model during VFL training. The poisoned model is furtherﬁne-tuned in the TGA phase to enhance its efﬁcacy in complexmulti-classiﬁcation tasks. Compared to the existing methods,the proposed TECB achieves a highly effective backdoor attackwith very limited information about the target class samples,which is more practical in typical VFL settings. Experimentalresults verify the superior performance of TECB, achieving above97% attack success rate (ASR) on three widely used datasets(CIFAR10, CIFAR100, and CINIC-10) with only 0.1% of targetlabels known, which outperforms the state-of-the-art attack meth-ods. This study uncovers the potential backdoor risks in VFL,enabling the development of secure VFL applications in areaslike ﬁnance, healthcare, and beyond.

随着深度学习算法在各个领域的成功应用，研究对抗性攻击以确保深度学习模型在实际应用中的安全性已成为一个重要的研究课题。后门攻击是深度网络上的一种对抗性攻击形式，攻击者向受害者提供投毒数据用于训练模型，然后在测试时通过展示特定的微小触发器模式来激活攻击。大多数最先进的后门攻击要么提供可能通过视觉检查识别出的错误标记的投毒数据，要么在投毒数据中暴露触发器，或者使用噪声来隐藏触发器。我们提出了一种新型的后门攻击方式，其中投毒数据看起来自然且标签正确，更重要的是，攻击者在投毒数据中隐藏了触发器，并在测试之前一直保密。我们对各种图像分类设置进行了广泛研究，并表明，尽管模型在干净数据上表现良好，但我们的攻击可以通过在未见过的图像上随机位置粘贴触发器来欺骗模型。我们还表明，使用最先进的后门攻击防御算法无法轻易防御我们提出的攻击。

8.With the success of deep learning algorithms in various do-mains, studying adversarial attacks to secure deep modelsin real world applications has become an important researchtopic. Backdoor attacks are a form of adversarial attacks ondeep networks where the attacker provides poisoned data tothe victim to train the model with, and then activates the at-tack by showing a speciﬁc small trigger pattern at the testtime. Most state-of-the-art backdoor attacks either providemislabeled poisoning data that is possible to identify by visualinspection, reveal the trigger in the poisoned data, or use noiseto hide the trigger. We propose a novel form of backdoor at-tack where poisoned data look natural with correct labels andalso more importantly, the attacker hides the trigger in thepoisoned data and keeps the trigger secret until the test time.We perform an extensive study on various image classiﬁca-tion settings and show that our attack can fool the model bypasting the trigger at random locations on unseen images al-though the model performs well on clean data. We also showthat our proposed attack cannot be easily defended using astate-of-the-art defense algorithm for backdoor attacks.

后门攻击对使用第三方数据进行人工智能开发的实践构成了重大威胁。数据可能被投毒，使得训练好的模型在预定义触发器模式出现时表现异常，从而为攻击者带来非法利益。虽然大多数提出的后门攻击都是脏标签攻击，但保持数据标签不变的干净标签攻击更能躲避人工检查，因此更为理想。然而，设计一个有效的干净标签攻击是一项具有挑战性的任务，且现有的干净标签攻击表现欠佳。在本文中，我们提出了一种新颖的机制来开发具有出色攻击性能的干净标签攻击。其关键组件是一个触发器模式生成器，它与一个替代模型以交替的方式进行训练。我们提出的机制具有灵活性和可定制性，允许为单个或多个目标标签设置不同类型的后门触发器和行为。我们的后门攻击可以达到接近完美的攻击成功率，并能绕过所有最先进的后门防御措施，这在标准基准数据集上的综合实验中得到了证明。

7.Backdoor attacks pose a critical concern to the practice of us-ing third-party data for AI development. The data can be poi-soned to make a trained model misbehave when a predefinedtrigger pattern appears, granting the attackers illegal benefits.While most proposed backdoor attacks are dirty-label, clean-label attacks are more desirable by keeping data labels un-changed to dodge human inspection. However, designing aworking clean-label attack is a challenging task, and exist-ing clean-label attacks show underwhelming performance. Inthis paper, we propose a novel mechanism to develop clean-label attacks with outstanding attack performance. The keycomponent is a trigger pattern generator, which is trained to-gether with a surrogate model in an alternating manner. Ourproposed mechanism is flexible and customizable, allowingdifferent backdoor trigger types and behaviors for either sin-gle or multiple target labels. Our backdoor attacks can reachnear-perfect attack success rates and bypass all state-of-the-art backdoor defenses, as illustrated via comprehensive ex-periments on standard benchmark datasets.

近年来，扩散模型在高质量图像生成领域取得了显著成功，受到了越来越多的关注。与此同时，人们对扩散模型相关的安全威胁也日益担忧，这主要是因为它们易受恶意利用。值得注意的是，最近的研究揭示了扩散模型对后门攻击的脆弱性，即通过相应的触发器生成特定的目标图像。然而，现有的后门攻击方法依赖于手工设计的触发器生成函数，这些函数通常表现为输入噪声中融入的可辨识模式，因此容易被人检测出来。在本文中，我们提出了一种创新且多功能的优化框架，旨在获取隐形触发器，从而增强插入后门的隐蔽性和鲁棒性。我们提出的框架既适用于无条件扩散模型，也适用于条件扩散模型，值得注意的是，我们是首次在文本引导的图像编辑和修复流程中演示了对扩散模型的后门攻击。此外，我们还展示了条件生成中的后门可以直接应用于模型水印，以进行模型所有权验证，这进一步提升了所提出框架的重要意义。在各种常用采样器和数据集上的大量实验验证了所提出框架的有效性和隐蔽性。

6.In recent years, diffusion models have achieved remarkable success in the realmof high-quality image generation, garnering increased attention. This surge ininterest is paralleled by a growing concern over the security threats associated withdiffusion models, largely attributed to their susceptibility to malicious exploitation.Notably, recent research has brought to light the vulnerability of diffusion modelsto backdoor attacks, enabling the generation of specific target images throughcorresponding triggers. However, prevailing backdoor attack methods rely on man-ually crafted trigger generation functions, often manifesting as discernible patternsincorporated into input noise, thus rendering them susceptible to human detection.In this paper, we present an innovative and versatile optimization framework de-signed to acquire invisible triggers, enhancing the stealthiness and resilience ofinserted backdoors. Our proposed framework is applicable to both unconditionaland conditional diffusion models, and notably, we are the pioneers in demon-strating the backdooring of diffusion models within the context of text-guidedimage editing and inpainting pipelines. Moreover, we also show that the backdoorsin the conditional generation can be directly applied to model watermarking formodel ownership verification, which further boosts the significance of the pro-posed framework. Extensive experiments on various commonly used samplers anddatasets verify the efficacy and stealthiness of the proposed framework.

尽管预训练语言模型得到了广泛应用，但已被证明易受后门攻击。后门攻击旨在通过向部分训练样本注入触发器和修改标签来在模型中引入针对性漏洞。传统的文本后门攻击存在几个缺陷：触发器导致自然语言表达异常，且被投毒的样本标签被错误标注。这些缺陷降低了攻击的隐蔽性，并且容易被防御模型检测到。在本研究中，我们介绍了Cbat，这是一种新颖且高效的方法，用于执行带文本风格的干净标签后门攻击，该方法不需要外部触发器，且被投毒的样本标签正确。具体来说，我们利用提示调优的强大少样本学习能力，开发了一个句子重写模型，以生成干净标签的被投毒样本。然后，Cbat通过被投毒的样本，将文本风格作为抽象触发器注入到受害模型中。我们还介绍了一种防御后门攻击的算法，名为CbatD，它通过定位最低训练损失并计算特征相关性，有效地擦除被投毒的样本。在文本分类任务上的实验表明，我们的Cbat和CbatD在文本后门攻击和防御方面均表现出整体竞争力。值得注意的是，Cbat在没有触发器的情况下，在干净标签后门攻击基准测试中取得了领先结果。

5.Despite being widely applied, pre-trained languagemodels have been proven vulnerable to backdoor attacks. Back-door attacks are designed to introduce targeted vulnerabilities intomodels by poisoning a subset of training samples through triggerinjection and label modiﬁcation. Traditional textual backdoor at-tacks suffer several ﬂaws: the triggers lead to abnormal naturallanguage expressions, and poisoned sample labels are mistakenlylabeled. These ﬂaws reduce the stealthiness of the attack and canbe easily detected by defense models. In this study, we introduceCbat, a novel and efﬁcient method to perform clean-label backdoorattack with text style, which does not require external trigger,and the poisoned samples are correctly labeled. Speciﬁcally, wedevelop a sentence rewriting model by leveraging the powerfulfew-shot learning capability of prompt tuning to generate cleanlabel poisoned samples. Cbat then injects text style as an abstracttrigger into the victim model through poisoned samples. We alsointroduce an algorithm for defending against backdoor attacks,named CbatD, which effectively erases the poisoned samples bylocating the lowest training loss and calculating feature relevance.The experiments on text classiﬁcation tasks demonstrate that ourCbat and CbatD show overall competitive performance in textualbackdoor attack and defense. It is noteworthy that Cbat attainedleading results in the clean-label backdoor attack benchmark with-out triggers.

后门攻击将恶意构建的数据注入机器学习模型的训练集中，以便在测试时，训练好的模型将带有后门触发器的输入错误分类为攻击者期望的输出。为了使后门攻击能够躲过人工检查，注入的数据必须看起来标签正确。具有这种特性的攻击通常被称为“干净标签攻击”。现有干净标签后门攻击的有效性严重依赖于对整个训练集的了解。然而，在实际操作中，获取这些知识成本高昂，甚至不可能，因为训练数据通常是从多个独立来源收集的（例如，来自不同用户的面部图像）。后门攻击是否仍然构成真实威胁，这个问题仍有待解答。在本文中，我们通过设计一种基于目标类样本和公开非分布内数据的知识的算法来发动干净标签后门攻击，对这个问题给出了肯定的回答。我们仅插入了恶意制作的示例，总数为目标类数据大小的0.5%和训练集大小的0.05%，就可以操纵在这个被投毒的数据集上训练的模型，使其在示例被注入后门触发器时，将任意类别的测试示例分类为目标类；同时，训练好的模型在没有触发器的情况下，对典型测试示例仍然保持良好的准确性，就像它是在干净的数据集上训练的一样。我们的攻击在不同数据集和模型上都非常有效，甚至在触发器被注入到物理世界中时也是如此。我们探索了防御措施的空间，并发现Narcissus（一种防御方法）在其原始形式或经过简单调整后，可以规避最新的最先进的防御措施。我们研究了这种令人费解的有效性的原因，并发现，由于我们的攻击合成的触发器包含与目标类的原始语义特征一样持久的耐用特征，任何试图移除这些触发器的行为都会不可避免地首先损害模型的准确性。

4.Backdoor attacks inject maliciously constructed data into a train-ing set of machine learning models so that, at test time, the trainedmodel misclassifies inputs patched with a backdoor trigger as anattacker-desired output. For backdoor attacks to bypass human in-spection, it is essential that the injected data appear to be correctlylabeled. The attacks with such property are often referred to as“clean-label attacks.” The effectiveness of existing clean-label back-door attacks crucially relies on knowledge about the entire trainingset. However, in practice, obtaining this knowledge is costly orimpossible as training data are often gathered from multiple inde-pendent sources (e.g., face images from different users). It remainsaquestion of whether backdoor attacks still present real threats.In this paper, we provide an affirmative answer to this question bydesigning an algorithm to mount clean-label backdoor attacks basedon the knowledge of samples from the target class and public out-of-distribution data. By inserting maliciously-crafted examples totalingjust 0.5% of the target-class data size and 0.05% of the training setsize, we can manipulate a model trained on this poisoned datasetto classify test examples from arbitrary classes into the target classwhen the examples are patched with a backdoor trigger; at the sametime, the trained model still maintains good accuracy on typicaltest examples without the trigger as if it were trained on a cleandataset. Our attack is highly effective across datasets and models,and even when the trigger is injected into the physical world.We explore the space of defenses and find that Narcissus canevade the latest state-of-the-art defenses in their vanilla form orafter a simple adaptation. We study the cause of the intriguingeffectiveness and find that because the trigger synthesized by ourattack contains durable features as persistent as the original seman-tic features of the target class, any attempt to remove such triggerswould inevitably hurt the model accuracy first.

近期研究表明，深度神经网络（DNN）易受后门攻击。被攻击的模型在良性样本上表现正常，但一旦出现攻击者指定的触发模式，其预测就会被误导。目前，干净标签后门攻击通常被认为是最隐蔽的方法，攻击者只能在不修改标签的情况下，对目标类的样本进行投毒。然而，这些攻击很难成功。在本文中，我们揭示了干净标签攻击之所以困难，主要在于投毒样本中包含的与目标类相关的“鲁棒特征”的对抗性影响。具体来说，鲁棒特征容易被受害模型学习，从而破坏了触发模式的学习。基于这些理解，我们提出了一种简单而有效的插件方法，通过投毒“困难”样本而非随机样本来增强干净标签后门攻击。我们采用了三种经典的难度指标作为示例来实现我们的方法。通过在大量基准数据集上的实验，我们证明了我们的方法能够一致地改进基础攻击。

3.Recent studies demonstrated that deep neural networks (DNNs) are vulnerable to backdoor attacks.The attacked model behaves normally on benign samples, while its predictions are misled wheneveradversary-speciﬁed trigger patterns appear. Currently, clean-label backdoor attacks are usually regardedas the most stealthy methods in which adversaries can only poison samples from the target class with-out modifying their labels. However, these attacks can hardly succeed. In this paper, we reveal that thediﬃculty of clean-label attacks mainly lies in the antagonistic effects of ‘robust features’ related to the tar-get class contained in poisoned samples. Speciﬁcally, robust features tend to be easily learned by victimmodels and thus undermine the learning of trigger patterns. Based on these understandings, we proposea simple yet effective plug-in method to enhance clean-label backdoor attacks by poisoning ‘hard’ insteadof random samples. We adopt three classical diﬃculty metrics as examples to implement our method. Wedemonstrate that our method can consistently improve vanilla attacks, based on extensive experimentson benchmark datasets.

深度神经网络（DNN）易受后门攻击的影响，攻击者通过操纵一小部分训练数据，使得受害模型对良性样本进行正常预测，但将被触发样本分类为目标类别。后门攻击是一种新兴且威胁巨大的训练阶段威胁，给基于DNN的应用带来了严重风险。在本文中，我们重新审视了现有后门攻击的触发模式。我们发现，这些模式要么可见，要么不够稀疏，因此隐蔽性不足。更重要的是，简单地结合现有方法来设计一个有效的稀疏且隐形的后门攻击是不可行的。为了解决这个问题，我们将触发器的生成表述为一个具有稀疏性和隐形性约束的双层优化问题，并提出了一种有效的方法来解决它。所提出的方法被称为稀疏且隐形的后门攻击（SIBA）。我们在不同设置下的基准数据集上进行了大量实验，验证了我们的攻击的有效性及其对现有后门防御措施的抵抗力。

2.Deep neural networks (DNNs) are vulnerable tobackdoor attacks, where the adversary manipulates a smallportion of training data such that the victim model predictsnormally on the benign samples but classifies the triggered samplesas the target class. The backdoor attack is an emerging yetthreatening training-phase threat, leading to serious risks inDNN-based applications. In this paper, we revisit the triggerpatterns of existing backdoor attacks. We reveal that theyare either visible or not sparse and therefore are not stealthyenough. More importantly, it is not feasible to simply combineexisting methods to design an effective sparse and invisiblebackdoor attack. To address this problem, we formulate the triggergeneration as a bi-level optimization problem with sparsity andinvisibility constraints and propose an effective method to solveit. The proposed method is dubbed sparse and invisible backdoorattack (SIBA). We conduct extensive experiments on benchmarkdatasets under different settings, which verify the effectivenessof our attack and its resistance to existing backdoor defenses

1.数据集蒸馏作为一种提高机器学习模型数据效率的技术，日益受到关注。该技术将大型数据集的知识封装到一个较小的合成数据集中。在这个较小的蒸馏数据集上训练的模型，其性能可以与在原始训练数据集上训练的模型相媲美。然而，现有的数据集蒸馏技术主要旨在在资源使用效率和模型效用之间达到最佳平衡，其带来的安全风险尚未被探索。本研究首次针对在图像领域由数据集蒸馏模型蒸馏的数据上训练的模型进行了后门攻击。具体而言，我们在蒸馏过程中向合成数据注入触发器，而不是在模型训练阶段（所有之前的攻击都是在这一阶段进行的）。我们提出了两种类型的后门攻击，即NAIVEATTACK和DOORPING。NAIVEATTACK简单地在初始蒸馏阶段向原始数据添加触发器，而DOORPING则在整个蒸馏过程中迭代更新触发器。我们对多个数据集、架构和数据集蒸馏技术进行了广泛评估。实证评估表明，在某些情况下，NAIVEATTACK取得了不错的攻击成功率（ASR）分数，而DOORPING在所有情况下都达到了更高的ASR分数（接近1.0）。此外，我们还进行了一项全面的消融研究，以分析可能影响攻击性能的因素。最后，我们评估了多种针对我们的后门攻击的防御机制，并表明我们的攻击实际上可以绕过这些防御机制。

Dataset distillation has emerged as a prominent techniqueto improve data efﬁciency when training machine learningmodels. It encapsulates the knowledge from a large datasetinto a smaller synthetic dataset. A model trained on thissmaller distilled dataset can attain comparable performanceto a model trained on the original training dataset. How-ever, the existing dataset distillation techniques mainly aimat achieving the best trade-off between resource usage efﬁ-ciency and model utility. The security risks stemming fromthem have not been explored. This study performs the ﬁrstbackdoor attack against the models trained on the data dis-tilled by dataset distillation models in the image domain.Concretely, we inject triggers into the synthetic data duringthe distillation procedure rather than during the model train-ing stage, where all previous attacks are performed. We pro-pose two types of backdoor attacks, namely NAIVEATTACKand DOORPING.NAIVEATTACK simply adds triggers to theraw data at the initial distillation phase, while DOORPINGiteratively updates the triggers during the entire distillationprocedure. We conduct extensive evaluations on multipledatasets, architectures, and dataset distillation techniques.Empirical evaluation shows that NAIVEATTACK achieves de-cent attack success rate (ASR)scores in some cases, whileDOORPING reaches higher ASR scores (close to 1.0) in allcases. Furthermore, we conduct a comprehensive ablationstudy to analyze the factors that may affect the attack perfor-mance. Finally, we evaluate multiple defense mechanismsagainst our backdoor attacks and show that our attacks canpractically circumvent these defense mechanisms