面对图像分类的对抗性攻击策略

史少杰 1024041004 软件工程 数据挖掘与智能软件

摘　要　深度神经网络 （DNN） 广泛应用于许多应用，并实现了最先进的性能。但是，DNN 在结构上缺乏对用户的透明度和可解释性。攻击者可以利用此功能在 DNN 结构中嵌入木马，例如在 DNN 中插入后门，使 DNN 可以同时学习正常的 main 任务和其他恶意任务。此外，DNN 依赖于数据集进行训练。攻击者可以篡改训练数据以干扰 DNN 训练过程，例如在输入数据上附加触发器。由于 DNN 结构和数据的缺陷，后门攻击可能对 DNN 的安全构成严重威胁。受后门攻击的 DNN 在良性输入上表现良好，同时在触发器附加的输入上输出攻击者指定的标签。后门攻击几乎可以在机器学习管道的每个阶段进行。尽管对图像分类的后门攻击有一些研究，但在该领域仍然很少见系统综述。本文对后门攻击进行了全面回顾。根据攻击者是否能够访问训练数据，将各种后门攻击分为两种：基于中毒的攻击和非基于中毒的攻击。我研究了时间线中每项工作的细节，讨论了它的贡献和不足。最后，设计实验验证了几种攻击在不同数据集以及不同中毒率下的表现，并给出相关结论。

关键词　图像识别；中毒攻击；网络安全；后门攻击；攻击策略

Adversarial Attack Strategies in Image Classification

SHI Shaojie　1024041004　Software engineering　Data mining and intelligent software

Abstract　Deep Neural Networks (DNNs) are widely applied in numerous applications and have achieved state-of-the-art performance. However, DNNs inherently lack transparency and interpretability for users. Attackers can exploit this characteristic to embed Trojans into DNN structures, such as inserting backdoors that enable DNNs to simultaneously learn the primary task and other malicious tasks. Moreover, DNNs rely on datasets for training, which attackers can tamper with to disrupt the training process, such as by attaching triggers to input data. Due to vulnerabilities in DNN structures and datasets, backdoor attacks pose significant threats to the security of DNNs. A DNN compromised by backdoor attacks performs well on benign inputs but outputs attacker-specified labels for inputs with attached triggers. Backdoor attacks can be conducted at nearly every stage of the machine learning pipeline. Although there has been some research on backdoor attacks in image classification, comprehensive reviews in this domain remain rare. This paper provides an in-depth review of backdoor attacks. Various backdoor attacks are classified into two categories based on whether the attacker has access to the training data: poisoning-based attacks and non-poisoning-based attacks. I have analyzed the details of each work along a timeline, discussing its contributions and limitations. Finally, experiments were designed to evaluate the performance of several attacks across different datasets and poisoning rates, and relevant conclusions are presented.

Key words　image recognition; poisoning attacks; cybersecurity; backdoor attacks; attack strategies

1 引言

近年来，深度学习取得了显著进步，推动了其在不同行业和学术领域的广泛采用。这种快速整合在医疗保健、教育、汽车和物流等行业中尤为明显，这些行业越来越多地利用深度学习来促进创新[1]。深度学习成功的一个关键因素是它能够从数据中提取复杂的模式。虽然此功能提供了显著的优势，但也带来了与可解释性相关的重大挑战。尽管深度

学习模型具有很强的预测性能，但通常缺乏透明度和可解释性，因此很难为具体预测提供明确的理由。深度学习模型的黑盒性质已被证明会使它们面临相当大的安全漏洞[2]。特别地，分类模型（例如用于图像识别的分类模型）已被证明容易受到操纵，多次对抗性攻击实例成功地破坏了他们的决策过程。例如，[3]的开创性工作强调了图像分类模型对对抗性示例的脆弱性，表明应用于输入图像的难以察觉的扰动会导致严重的错误分类。标志性的“Panda-Gibbon”图像是使用[4]中提出的生成对抗性扰动的方法创建的，它突出地说明了深度学习模型固有的脆弱性，尽管它们很复杂。从那时起，已经确定了其他几种对抗性威胁，影响了广泛的学习任务[5]。

在实际场景中，后门攻击是一种重大威胁，尤其是在分类输出驱动自动化决策过程的情况下[6]。后门攻击有意在输入空间内的虚假特征（称为触发器）与特定分类结果之间建立关系。一旦建立了这种关联，受损模型就会对干净的图像（即未更改的图像）和后门图像（即包含触发器的图像）进行不同的分类。因此，后门攻击通过插入不需要的行为（称为后门任务）来破坏模型决策的完整性，而不会影响其正确执行识别干净图像的原始分类任务的能力[7]。在图 1 中，我展示了一个干净的图像和一个后门版本，以及一个成功被攻破（后门）的模型的分类结果。这种后门攻击实际上可以通过简单地在停车标志上贴上黄色贴纸来执行。在自动驾驶环境中，此类攻击可能会产生严重的安全隐患。

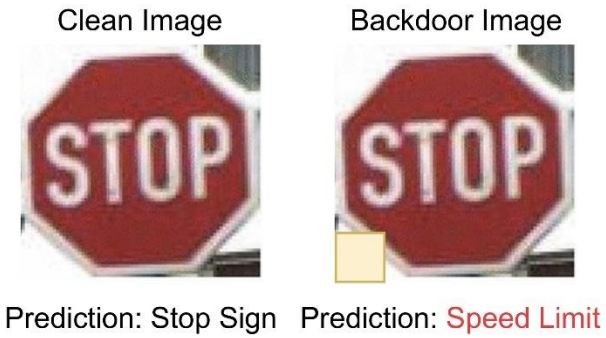

图 1　后门图像（右）及其相应的干净图像（左）示例。作为触发器的黄色方块已添加到后门图像的左下角。

要成功对模型执行后门攻击，攻击者必须破坏目标模型的训练管道。这使得当模型训练外包给第三方时，例如通过基于云的机器学习即服务平台[8]，后门攻击特别危险。在这种情况下，攻击者可以操纵训练数据或程序，在受害者不知情的情况下注入后门任务。在深度学习社区中广泛使用预训练模型权重加剧了这种威胁。例如，最近的一项行业调查[9]显示，48.1% 的参与者使用第三方权重进行模型训练，进一步放大了后门攻击的潜在风险。

为了应对当前深度学习方法难以解释的性质所带来的挑战，机器学习安全已成为一个关键的研究领域[10]。虽然它不能直接解决可解释性问题，但机器学习安全寻求开发新方法来抵御已知威胁，确保深度学习可以安全部署。在该领域，一项新兴的工作专注于开发专门用于对抗后门攻击的防御策略。这些策略旨在通过从模型中删除后门任务来降低风险，同时保留其对干净输入进行分类的能力。然而，尽管该领域取得了重大进展，但由于实际限制，当前提案在不同环境中的一致性和可靠性仍不确定。许多提出的方法都使用有限范围的攻击、数据集、模型架构和数据可用性条件进行评估。这引发了人们对它们在各种现实场景中的泛化性和有效性的质疑。

2 攻击策略

在本节中，我对当前最先进的针对图像分类神经网络的对抗性攻击进行了综述。我将这些攻击分为五类：白盒攻击、黑盒攻击、中毒攻击、提取攻击和推理攻击。本文重点深入剖析这些攻击的工作原理，同时对其性能进行评估。下文对所评述的攻击进行了快速概览，并在攻击效果、可迁移性以及执行时间和扰动大小等性能指标方面对它们进行了比较。需要注意的是，大多数被评述的攻击集中于通过生成扰动来破坏深度神经网络（DNN）。这是因为在干净图像中加入扰动是一种有效的攻击方法，生成的对抗样本几乎与原始图像无异（见图 9），因此引起了研究界的广泛关注。

2.1　白盒攻击

2.1.1　L-BFGS

对抗样本最早由 Szegedy 等人提出[11]。他们发现，在图像 x 上添加一个小的扰ρ，可以生成一个对抗样本，从而成功欺骗深度学习模型。为计算适当的扰动大小，作者尝试解决以下优化问题：

$$min\ |\rho|_2\ subject\ to\ f(x\ +\ \rho)\ =\ l, x\ +\ \rho \in [0,1]^m$$

然而，由于上述问题难以直接求解，作者采用了带约束的 L-BFGS 方法（Box-Constrained L-BFGS）[12]来估算解，如下式所示：

$$\min c \cdot |\rho| + L_f(x + \rho, l) \; subject \; to$$
$$x + \rho \in [0,1]^m$$

这一方法通过寻找满足条件 f(x+ρ)=l 的最小值，同时计算分类器的损失函数。

作者观察到，由带约束的 L-BFGS 生成的对抗样本与原始图像几乎完全一致（即扰动不可察觉）。他们还注意到，这些对抗样本能够成功欺骗其他深度神经网络模型（即具有迁移性）。这一研究结果引发了人们对深度学习系统安全性的担忧，并激发了对抗性机器学习研究的广泛兴趣。

2.1.2 快速梯度方法

快速梯度方法 FGSM（Fast Gradient Sign Method）[13]被提出为一种高效算法，可为任意给定图像生成扰动。与 L-BFGS[14]相比，FGSM 具有以下两个显著区别：（1）其对抗样本采用 L∞度量；（2）其目的是提供一种快速生成对抗样本的方法，这些样本可以用于对抗训练，从而提高深度神经网络（DNN）模型抵御对抗攻击的鲁棒性。

更正式地说，给定图像 x，FGSM 使用以下公式计算扰动：

$$\{x\} = x - \epsilon \cdot sign\left(\nabla_x J(\theta, x, y)\right)$$

其中，$\nabla_x J(\cdot)$表示针对给定神经网络的损失函数（如分类交叉熵）关于输入 x 的梯度，ϵ是用于约束扰动大小的小常数。换句话说，FGSM 根据梯度的符号，通过对 x 的每个像素加上或减去ϵ来生成对抗样本 x。

FGSM 能够生成有效的对抗样本，成功欺骗不同的 DNN 模型。然而，从文献[15]中的实验结果可以看出，FGSM 需要添加较大的扰动量来生成强大的对抗样本，这可能导致图像分辨率的明显失真。

2.2 黑盒攻击

2.2.1 边界攻击

边界攻击（Boundary Attack）[16]通过在图像集合的决策边界上进行拒绝采样来实施攻击。该方法旨在寻找使目标图像被错误分类所需的最小扰动，并通过逐步靠近原始输入的超平面来实现。

边界攻击的初始阶段设置了一个易于检测的大扰动，并随机沿决策边界向目标类别移动，从而有效优化最初的扰动。此攻击包含两个关键参数：扰动的长度ρ和朝向初始图像的步δ。这两个参数都会根据边界的局部几何结构进行调整。如果扰动是对抗性的（即使目标图像被错误分类），则会向原始输入图像迈出一个小步。当算法逐渐接近输入图像时，决策边界变得更加平坦，步长δ必须减小以继续推进。攻击在δ收敛至零时结束。

边界攻击是一种针对深度学习模型的强大方法，其效果优于基于梯度的白盒攻击方法（如 FGSM[17] 和 DeepFool[18]）。实验表明，该方法能够有效破坏训练于 MNIST[19]、CIFAR-10[20]、VGG-19[21]、ResNet-50[22] 和 ImageNet[23]数据集上的模型。

2.2.2 零阶优化攻击

零阶优化（Zeroth-Order Optimization, ZOO）攻击[24]利用零阶随机坐标下降（zeroth-order stochastic coordinate descent）生成扰动。作者将 Carlini & Wagner 攻击[25]适配至黑盒威胁模型，通过修改损失函数并近似梯度实现了这一点。

具体来说，作者提出了一种新的损失函数 f(x, t)，该函数仅依赖模型输出和目标类别标签。攻击通过零阶优化对这一损失函数进行优化。此外，ZOO 使用随机坐标下降来近似模型的梯度，而不是传统的反向传播方法。零阶随机坐标下降用于针对模型进行攻击并提取有关梯度的信息。在特殊情况下，还会采用降维、分层攻击和重要性采样等技术来优化损失函数。

实验表明，ZOO 攻击极其有效。在非目标攻击场景下，该方法在 MNIST[26]和 CIFAR-10[27]数据集上的成功率达到 100%。在目标攻击场景下，ZOO 攻击对 MNIST 的成功率为 98.9%，对 CIFAR-10 的成功率为 97%。

2.2.3 空间变换攻击

顾名思义，空间变换攻击[28]通过对输入图像进行平移和旋转来生成对抗样本。为了确保与干净图像的视觉相似性，扰动范围被限制为最多 30° 的旋转和每个方向最多 10% 的平移。

最优扰动是通过超参数优化（通常称为网格搜索）计算得出的。网格搜索是一种广泛的搜索方法，用于在超参数空间的子集内寻找最优参数。在本例中，它用于为给定模型确定最优的扰动参数组合。计算得到的旋转和平移参数组合会应用于整个输入图像组。某种意义上，空间变换攻击所发现的扰动是通用的。

实验表明，空间变换攻击能够取得显著的效

果，成功欺骗训练于 MNIST[29]、CIFAR-10[30]和 ImageNet[31]数据集的多种深度学习模型。

2.3 中毒攻击

2.3.1 对支持向量机的中毒攻击

如果攻击者完全访问模型的学习算法和训练数据，他们可以有效地创建中毒攻击，从而成功地针对支持向量机（SVM）进行攻击[32]。该算法可以进行核化，但完全依赖于输入空间中点的梯度。攻击者利用迭代的梯度上升方法来优化模型的非凸目标函数。与梯度下降不同，梯度上升采取的步骤与正梯度成比例（而不是负梯度），从而接近局部最大值，而不是局部最小值。

攻击者通过一个向量启动攻击，该向量复制目标类别中的一个随机点，并更改其分类标签。在实际操作中，任何深度足够接近对抗类边界的点都可以用来启动攻击。然后，使用梯度上升算法找到验证误差的梯度。在更新过程中，必须保持用于训练 SVM 分类器的数据集架构。通常在使用梯度上升时，线性搜索算法会用于找到最优解。在这种情况下，需要较大的步长来确保训练集的安全，而线性搜索则会非常耗费计算资源。为了解决这个问题，攻击将步长固定为一个常数值。每次更新后，都会重新计算最优解。

当验证误差低于某个阈值时，攻击结束。该攻击可以显著提高分类错误率，从初始的 2-5% 增加到 15-20%，即使只使用一个对抗性数据点。

2.3.2 对抗性后门嵌入攻击

一般来说，像激活聚类（activation clustering）[33]这样的后门检测算法对于大多数后门攻击是有效的。然而，它们未能考虑到更为强大的对抗性模型。对抗性后门嵌入攻击[34]通过在目标训练函数中加入一个次级损失函数来利用这些弱防御。次级损失函数作为惩罚项，当模型检测到扰动图像和非扰动图像之间的差异时，就会惩罚模型。双重目标函数使得攻击者能够在不影响模型分类准确率的情况下，设置一些约束条件，从而削弱模型的防御能力。

随着对抗性训练的收敛，后门输入和干净输入的分布也会收敛——最小化防御系统用于检测中毒攻击的差异。

2.3.3 输入模型共优化攻击

尽管对抗性样本和中毒模型之间存在差异，但这两种威胁模型具有相同的目标，即攻击神经网络并错误分类输入数据。输入模型共优化攻击（IMC）[35]旨在将这两种威胁模型统一起来。作者定义了一个统一的框架，赋予攻击者自由，可以选择生成对抗性样本或中毒模型。该攻击为数据集中的每个输入 x 生成对抗性样本 x。然后，扰动图像会被中毒模型错误分类为特定目标类别。IMC 攻击通过在模型和输入扰动之间来回迭代，直到攻击收敛，从而找到最优的对抗性样本和中毒模型。

值得注意的是，IMC 攻击可以通过调整基础算法以适应不同的约束，来适应各种攻击场景，其中一个例子就是 TrojanNN[36]攻击。

2.4 提取攻击

2.4.1 复制猫网络 (Copycat Networks)

Correia-Silva 等人[37]提出了一种方法，旨在通过仅使用干净图像查询网络，将目标网络复制成一个复制猫版本。复制猫方法经历两个阶段：生成虚假信息和训练复制猫网络。

攻击者首先通过选择大量图像来生成一个虚假的数据集，这些图像可以来源于目标网络，或者来自与目标无关的图像集。攻击者根据对目标模型的访问权限来选择数据集收集方式。原始图像标签对于攻击者是无用的，因此会从所有图像集中过滤掉。该阶段攻击者的主要目标是观察目标网络如何对图像进行分类。实现方法是将新生成的数据集输入目标模型，让其对所有图像进行分类。新生成的图像标签被称为“被窃取标签”。攻击者的目的是捕捉分类器的微小瑕疵，这样可以训练另一个网络，在相同的数据集上产生与目标模型相似的结果。

在生成虚假数据集和被窃取标签后，攻击者训练一个复制猫网络，模仿原始网络的定义过程：首先，攻击者为复制猫网络选择一个模型架构，而无需了解目标网络的架构。然后，攻击者根据目标模型的问题领域调整所选的模型架构。例如，攻击者可以改变其网络的输出数量，以匹配目标模型中的类别数量。理想情况下，模型是预训练的并且具有随机化的权重，但这不是必须的。生成复制猫网络的最后一步是使用虚假数据集和被窃取标签对生成的模型进行微调。这使得攻击者能够模拟目标模型的原始条件，并生成最有效的对抗性样本，从而攻击目标模型。

2.4.2 功能等效提取

模型提取攻击直接针对给定模型的最机密部分——其架构和参数。通过模型提取，之前在黑盒威胁模型下操作的攻击者能够有效地获取模型的白盒访问权限，具体方法是提取一个与目标模型完全相同的副本。模型提取是最难实现的对抗性目标之一，因为攻击者试图在仅能访问输入和输出的情况下生成模型的副本。功能等效提取[38]旨在构建一个神谕模型 O′，使得：

$$\forall x \in X, O'(x) = O(x)$$

功能等效提取方法适用于使 ReLU 激活函数的神经网络。该算法分为四个步骤。首先，关键点搜索 确定网络的输入，使得一个 ReLU 单元位于关键点。通过对两个值进行采样并将其输入到线性函数中，可以计算输入向量的斜率和截距，然后计算两向量的交点。如果有超过两个线性因子，则很难使真实值与预测值匹配。其次，构建复制神谕的下一步是权重恢复。为了形成权重矩阵 A(0)，他们计算神谕 O 在每个输入方向上的二阶导数。二阶导数用于计算相邻线性区域之间的差异。这个过程重复进行，直到整个矩阵 A(0)完成。第三，算法确定每个行向量 $A(0)_j$ 的符号，使用有关矩阵的全局信息。最后，使用最小二乘法近似神经网络隐藏层的架构。

在对 MNIST[39]进行测试时，功能等效提取方法生成的神谕准确率达到 100%，并且只有在超 100,000 个参数时才开始下降。对 CIFAR-10[40]进行测试时，准确率在超过 200,000 个参数后低于 100%。该方法的主要问题是不能扩展到其他更深层的神经网络，仅能在两层模型上有效工作。

3 实验设计

3.1 数据集和模型

对于 4 个常用的数据集(CIFAR-1、CIFAR-10、GTSRB 、 Tiny ImageNe) 和 5 个骨干模型(PreAct-ResNet18、VGG-19(无批处理规范层)、EfficientNet-B3 、 MobileNetV3-Large 、 DenseNet-161)，为了公平地衡量攻防方法对每个模型的性能影响，只对每个模型使用了基本版本的训练，而没有添加任何其他训练技巧（如增强）。表中总结了数据集的详情和普通训练的清洁准确率。表 1 总结了正常训练的数据集和准确率。

3.2 攻击和防御

我在每种情况下针对 2 种防御评估 2 种攻击，以及一种无防御的攻击。因此，共有 4 对评估。我在 2 个数据集和 2 个模型的基础上，考虑了 5 种中毒率，即每对中毒率分别为 0.1%、0.5%、 1%、 5%、10%， 因此总共有 4*2*2*5=80 对评估。每个模型的性能都通过指标来衡量，即 C-Acc、 ASR 和 R-Acc。

数据集	分类	训练/测试大小	图像大小	Clean Accuracy				
				PreAct-ResNet18	VGG-19	EfficientNet-B3	MobileNetV3-Large	DenseNet-161
CIFAR-10	10	50,000/10,000	32 ×32	93.90%	91.38%	64.69%	84.44%	86.82%
CIFAR-100	100	50,000/10,000	64 ×64	70.51%	60.21%	48.92%	50.73%	57.57%
GTSRB	43	39,209/12,630	32 ×32	98.46%	95.84%	87.39%	93.99%	92.49%
Tiny ImageNet	200	100,000/10,000	64 ×64	57.28%	46.13%	41.08%	38.78%	51.73%

表 1 数据集细节以及正常训练的干净准确率

4 结论

我首先在图 2 中展示了在一种模型结构（即 PreAct-ResNet18）和一种投毒率（即 5%）下，各种攻的性能分布。在上排，性能通过干净样本准确率（C-Acc）和攻击成功率（ASR）进行衡量。从攻击者的角度来看，理想的性能应同时具有较高的 C-Acc 和 ASR，即位于图的右上角。从防御者的角度来看，理想的性能应同时具有较高的 C-Acc 和较低的 ASR，即位于图的左上角。可以观察到，大多数颜色模式分布在相似的水平线上，这表明大多数防御方法可以在不显著降低干净样本准确率的情况下减轻后门攻击的影响。

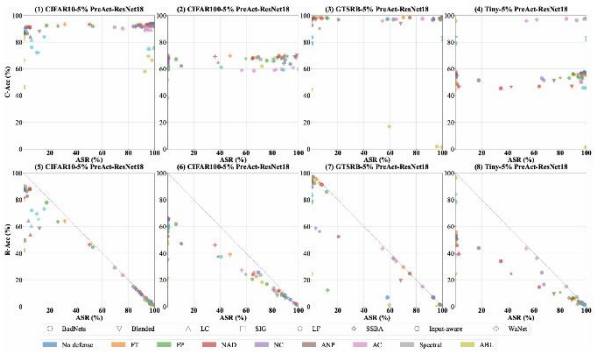

图 2 不同攻击的性能分布

在下排，性能通过鲁棒准确率（R-Acc）和 ASR

来衡量。通常情况下，ASR + R-Acc ≤ 1。从防御者的角度来看，理想的情况是 ASR 的降低值等于 R-Acc 的增加值，也就是说，在防御之后，被投毒样本的预测可以恢复到正确的类别。有趣的是，大多数颜色模式接近反对角线（即 ASR + R-Acc ≈ 1），但在 CIFAR-100（第二列）和 Tiny ImageNet（最后一列）中存在一些偏离反对角线的现象。我认为这与数据集的类别数量高度相关。类别数量较多时，在防御后恢复正确预测变得更加困难。

总之，上述分析表明了一个有趣的问题：中毒率越高的攻击并不意味着攻击性能越好，它可能更容易被某些防御方法抵御。原因在于，较高的中毒率会突出中毒样本和干净样本之间的差异，而这将被自适应防御所利用。这一点引发了两个值得在未来进一步探讨的有趣问题:如何利用更少的中毒样本达到理想的攻击性能，以及如何防御低中毒率的弱攻击。此外，考虑到权重初始化和某些方法机制的随机性，多次重复上述评估之后，虽然出现了一些波动，但 ASR 曲线的趋势与图 2 类似。

参 考 文 献

[1] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, “A survey on deep learning: Algorithms, techniques, and applications,” ACM Computing Surveys (CSUR), vol. 51, no. 5, pp. 1–36, 2018.

[2] X. Wang, J. Li, X. Kuang, Y.-a. Tan, and J. Li, “The security of machine learning in an adversarial setting: A survey,” Journal of Parallel and Distributed Computing, vol. 130, pp. 12–23, 2019.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[4] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017.

[5] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, “Backdoor learning: A survey,” IEEE Transactions on Neural Networks and Learning Systems, 2022.

[6] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, “Abs: Scanning neural networks for back-doors by artificial brain stimulation,” in Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 1265–1282.

[7] K. Grosse, L. Bieringer, T. R. Besold, and A. Alahi, “Towards more practical threat models in artificial intelligence security,” arXiv preprint arXiv:2311.09994, 2023.

[8] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” IEEE transactions on neural networks and learning systems, vol. 30, no. 9, pp. 2805–2824, 2019.

[9] X. Sheng, Z. Han, P. Li, and X. Chang, “A survey on backdoor attack and defense in natural language processing,” in 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS). IEEE, 2022, pp. 809–820.

[10] S.-H. Chan, Y. Dong, J. Zhu, X. Zhang, and J. Zhou, “Baddet: Backdoor attacks on object detection,” in European Conference on Computer Vision. Springer, 2022, pp. 396–412.

[11] Y. Li, Y. Li, Y. Lv, Y. Jiang, and S.-T. Xia, “Hidden back door attack against semantic segmentation models,” arXiv preprint arXiv:2103.04038, 2021.

[12] P. Cheng, Z. Wu, W. Du, and G. Liu, “Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review,” arXiv preprint arXiv:2309.06055, 2023.

[13] S. Zhao, M. Jia, Z. Guo, L. Gan, J. Fu, Y. Feng, F. Pan, and L. A. Tuan, “A survey of backdoor attacks and defenses on large language models: Implications for security measures,” arXiv preprint arXiv:2406.06852, 2024.

[14] Q. Le Roux, E. Bourbao, Y. Teglia, and K. Kallas, “A comprehensive survey on backdoor attacks and their defenses in face recognition systems,” IEEE Access, 2024.

[15] B. Yan, J. Lan, and Z. Yan, “Backdoor attacks against voice recognition systems: A survey,” arXiv preprint arXiv:2307.13643, 2023.

[16] Y. Wan, Y. Qu, W. Ni, Y. Xiang, L. Gao, and E. Hossain, “Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey,” IEEE Communications Surveys & Tutorials, 2024.

[17] B. Wu, H. Chen, M. Zhang, Z. Zhu, S. Wei, D. Yuan, and C. Shen, “Backdoorbench: A comprehensive benchmark of backdoor learning,” Advances in Neural Information Processing Systems, vol. 35, pp. 10546–10559, 2022.

[18] M. Zhu, S. Wei, H. Zha, and B. Wu, “Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features,” Advances in Neural Information Processing Systems, vol. 36, 2024.

[19] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia, “Adversarial unlearning of backdoors via implicit hypergradient,” in International Conference on Learning Representations, 2022.

[20] P. Zhao, P.-Y. Chen, P. Das, K. N. Ramamurthy, and X. Lin, “Bridging mode connectivity in loss landscapes and adversarial robustness,” arXiv preprint arXiv:2005.00060, 2020.

[21] K. Doan, Y. Lao, W. Zhao, and P. Li, “Lira: Learnable, imperceptible and robust backdoor attacks,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11966–11976.

[22] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Targeted backdoor attacks on deep learning systems using data poisoning,” arXiv preprint arXiv:1712.05526, 2017.

[23] M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns

by training set corruption without label poisoning," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 101–105.

[24] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, "Rethinking the back door attacks' triggers: A frequency perspective," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 16473–16481.

[25] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 16463–16472.

[26] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," Advances in Neural Information Processing Systems, vol. 33, pp. 34543464, 2020.

[27] Z. Wang, J. Zhai, and S. Ma, "Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15074–15084.

[28] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," arXiv preprint arXiv:2102.10369, 2021.

[29] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019, pp. 707–723.

[30] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, 20 "Strip: A defence against trojan attacks on deep neural networks," in Proceedings of the 35th annual computer security applications conference, 2019, pp. 113–125.

[31] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in International sympo sium on research in attacks, intrusions, and defenses. Springer, 2018, pp. 273–294.

[32] R. Zheng, R. Tang, J. Li, and L. Liu, "Pre-activation distributions expose backdoor neurons," Advances in Neural Information Processing Systems, vol. 35, pp. 18667–18680, 2022.

[33] ——, "Data-free backdoor removal based on channel lipschitzness," in European Conference on Computer Vision. Springer, 2022, pp. 175 191.

[34] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," Advances in Neural Information Processing Systems, vol. 34, pp. 16913–16925, 2021.

[35] S. Chai and J. Chen, "One-shot neural backdoor erasing via adversarial weight masking," Advances in Neural Information Processing Systems, vol. 35, pp. 22285–22299, 2022.

[36] Y. Li, X. Lyu, X. Ma, N. Koren, L. Lyu, B. Li, and Y.-G. Jiang, "Reconstructive neuron pruning for backdoor defense," in International Conference on Machine Learning. PMLR, 2023, pp. 19837–19854.

[37] H. Wang, Z. Xiang, D. J. Miller, and G. Kesidis, "Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic," in 2024 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2023, pp. 15–15.

[38] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 707–723.

[39] X. Qiao, Y. Yang, and H. Li, "Defending neural backdoors via genera tive distribution modeling," Advances in neural information processing systems, vol. 32, 2019.

[40] Y. Liu, M. Fan, C. Chen, X. Liu, Z. Ma, L. Wang, and J. Ma, "Backdoor defense with machine unlearning," in IEEE INFOCOM 2022-IEEE conference on computer communications. IEEE, 2022, pp. 280–289.