DOI:10.11918/202212023

第一类投毒攻击及其特征分析

王鹏博,陈思哲,黄晓霖

(上海交通大学 图像处理与模式识别研究所,上海 200240)

摘 要:为研究神经网络在面对安全威胁时的鲁棒性与可信性问题,聚焦于其在投毒攻击下的脆弱性,在系统分析第一类对抗攻击与第二类对抗攻击特征的基础上,结合神经网络在特征学习中的结构性缺陷,提出第一类投毒攻击的概念。通过理论分析建模,明确第一类投毒攻击与现有的"干净标签"、特征碰撞等投毒攻击在特征层面的本质差异。基于监督变分自编码器构建第一类投毒样本生成框架,并在ResNet50、VGG16、MobileNetV2等常用深度神经网络模型上开展实验。结果表明:第一类投毒攻击方法在不破坏标签一致性的前提下,有效干扰模型的分类决策,能够在典型神经网络构架上诱导模型产生分类错误。此外,防御实验表明:第一类投毒攻击可绕过现有主流防御机制,使现有主要防御机制失效。第一类投毒攻击具有较强的隐蔽性和破坏性,是一种值得深入研究的新型安全威胁形式,该攻击方法的提出对于未来构建更安全、鲁棒性更强的神经网络系统具有重要意义。

关键词:神经网络;投毒攻击;第一类错误;特征分析;稳健性

中图分类号:TP183 **文献标志码**:A **文章编号**:0367-6234(2025)09-0021-08

Type I poisoning attack and its feature analysis

WANG Pengbo, CHEN Sizhe, HUANG Xiaolin

(Institute of Pattern Analysis and Machine Intelligence, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: To investigate the robustness and trustworthiness of neural networks under security threats, this study focuses on their vulnerability to poisoning attacks. Based on a systematic analysis of the characteristics of type I and type II adversarial attacks, and in light of the structural deficiencies in neural network feature learning, the concept of type I poisoning attack is proposed. Theoretical modeling and analysis demonstrate fundamental feature-level distinctions between type I poisoning attacks and existing methods, such as "clean-label" or feature collision poisoning. A type I poisoned sample generation framework is built based on supervised variational autoencoders, and experiments on widely-used deep neural network architectures including ResNet50, VGG16, and MobileNetV2 are conducted. Results demonstrate that the proposed type I poisoning method effectively disrupts model classification decisions while preserving label consistency, successfully inducing misclassification across typical neural network architectures. Moreover, the defense experiments reveal that type I poisoning attacks can bypass existing mainstream defense mechanisms, rendering current primary countermeasures ineffective. With its strong stealth and disruptive capabilities, type I poisoning represents a novel security threat worthy of in-depth investigation. The development of this attack methodology holds significant implications for building more secure and robust neural network systems in the future.

Keywords: neural network; poisoning attack; type I error; feature analysis; robustness

以深度学习为代表的新一代人工智能技术在近年来得到了快速发展,在计算机视觉[1]、自然语言处理[2]、自动驾驶[3]、推荐系统[4]、医疗系统[5]等领域得到了广泛应用。在许多实际应用中,深度学习识别速度和精度已经超过传统算法以及人工识别。然而,神经网络在稳健性方面的不足所引发的安全性问题,已成为制约相关算法大规模应用的重要障碍,并引起了学术界的广泛关注与深入研究[6]。

目前,针对深度神经网络的攻击主要分为对抗攻击[7]和投毒攻击[8]。对抗攻击发生在模型训练完成后的部署阶段,攻击者在不改变目标深度学习模型的情况下,构造特定输入样本欺骗目标系统。投毒攻击发生在模型训练阶段,攻击者在训练数据中加入精心构造的异常数据,使模型产生特定的预

收稿日期:2022-12-08;**录用日期**:2023-02-18;**网络首发日期**:2025-07-04
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20250704.1520.002
基金项目:科技部重点研发计划(2023YFF1104202);国家自然科学基金(62376155)
作者简介:王鹏博(1997—),男,硕士研究生;黄晓霖(1984—),男,教授,博士生导师
通信作者:黄晓霖,xiaolinhuang@sjtu.edu.cn

测错误[9]。

早期的对抗攻击和投毒攻击主要瞄准神经网络的“过灵敏”缺陷,即在样本上施加微小变化,诱导神经网络出现分类错误。在投毒攻击领域中,攻击者通常会微小地篡改少量训练数据并赋予其正确标签,在模型训练过程中,对其进行破坏或设置可以通过特定方式触发的后门[10-12]。通过上述方式构造的投毒样本能够在不显著改变数据分布的情况下完成对模型的定向破坏,攻击者无需直接控制模型结构,即可在测试阶段引发期望的错误预测。该类攻击正是基于模型对非判别性微扰高度敏感的弱点展开的,体现出神经网络在特征学习过程中的鲁棒性缺陷。与“过灵敏”缺陷相反,Tang 等[13]在第一类对抗攻击研究中发现神经网络也存在“过懒惰”缺陷,即构造特定巨大差异的设计样本,使得神经网络保持不变,最终对样本识别失败。上述两种缺陷分别对应统计学中的第一类错误和第二类错误,二者在理论上有本质不同。

本文在 Tang 等[13]前期研究基础之上,从统计学意义上的第一类错误与第二类错误出发,将传统投毒攻击的概念进行系统拓展与精细化分类,提出了第一类投毒攻击。在此基础上,对两类投毒攻击在攻击目标、行为机制、标签一致性、特征干扰方式等方面进行深入的理论分析,明确二者在特征空间层面的本质差异。针对第一类投毒攻击,构建具体实现方法,并在多个深度神经网络模型上进行实验验证。通过在标准测试数据集上的效能评估,揭示该新型攻击方式的隐蔽性与破坏性,进一步指出现有防御机制在应对新型投毒攻击时的局限性。新形式的投毒攻击的提出,有助于完善攻击领域的类型划分,也为未来神经网络安全性的系统评估与多类型攻击防御策略的设计提供理论基础和实验支撑。

1 投毒攻击主要相关研究

投毒攻击是一种在模型训练阶段有意注入恶意样本以干扰模型学习过程,并在推理阶段诱导其产生错误预测的攻击方式。Stutz 等[14]、Goodfellow 等[15]提出了主要对模型训练过程施加控制的早期方法,但该方法实际运用难度较大。为了提高算法的可用性,Liu 等[10]提出了通过污染训练样本来实现模型干扰,该策略在学术界引发了广泛关注。Biggio 等[9]指出直接篡改训练数据的真实标签是一种具有代表性的投毒攻击手段,该手段能够有效干扰模型的学习过程,但由于其易被检测,隐蔽性较差,导致其在实际场景中的适用性受到了限制。

为克服以上问题,Shafahi 等[11]提出了一种将目标图像信息轻度叠加于正常图像上的投毒方法。该方法将目标图像和正常图像在每个像素上进行加权平均,生成具有攻击性的投毒样本。从数据检查者看来,图像并没有发生太大的变化,故难以将其移出训练集。训练模型时,神经网络会学习到投毒图像上叠加的两张图像的特征,从而错误地认为这些特征属于同一个类别。训练完成后,模型在推理阶段会将目标图像的特征误判为与投毒样本相同的标签,从而实现攻击者预期的错误分类。Turner 等[12]提出了另外一种思路,即通过在模型分类边界样本上添加触发器图案的方式向模型注入后门。在训练过程中,分类器会共同学习触发图案的特征和正常图像的特征,将其作为分类的依据。通过此方式训练的模型,会被注入一个带有触发器的后门,所有带此触发器的图像均会被判别为指定的类别。由于攻击过程中没有修改标签,添加的触发器也很难被人眼察觉,故此方法的隐蔽性很高。

近年来,Tang 等[13]、Carlini 等[16]提出了第一类对抗攻击的概念,这是一种针对神经网络的新型攻击方式。该对抗攻击会使对抗样本发生显著变化,但神经网络对其输出不变。在第一类对抗攻击过程中,攻击者利用分类模型对图像某些重要特征的忽略,通过引入显著可感知的扰动,使对抗样本在人眼看来发生了较大变化,但神经网络仍将其误判为原始类别。虽然这种方法已经在对抗攻击领域有了一定的应用,但经过充分的文献调研得知,在学术界尚未发现将第一类对抗攻击中的方法逆向应用于投毒攻击领域的研究。

2 两类投毒方法理论分析

2.1 投毒攻击分类

Madry 等[17]认为投毒攻击与对抗攻击有着深刻的联系,本文将现有的投毒方法定义为第二类投毒攻击,将第一类对抗攻击中神经网络“过懒惰”缺陷方法的逆应用定义为第一类投毒攻击,并通过理论分析和仿真验证分类的必要性和可行性。

基于投毒攻击和对抗攻击之间的紧密耦合,并结合统计学中第一类错误的定义,第一类投毒攻击的直观意义是通过大幅度篡改少量训练数据,使神经网络在正确标签的指引下,无法完整学习正确的图像分布,从而对特定图像产生极高的误判率。也就是说,即使投毒样本在外观上与原始样本存在显著差异,第一类投毒攻击仍能够诱导神经网络将其误判为目标图像所属类别。与第二类投毒攻击[10-12]相比,第一类投毒攻击将诱导神经网络学习缺失特征,而第二类投毒攻击则诱导神经网络学习

冗余特征。表 1 为深度学习模型攻击方法分类。目前学术界对第一类对抗攻击、第二类对抗攻击和第二类投毒攻击进行了大量深入的研究，但鲜少有针对第一类投毒攻击的研究。

表 1　现有针对深度学习模型的攻击方法分类

Tab. 1　Classification of existing attack methods for deep neural network

攻击类型	第一类攻击	第二类攻击
对抗攻击	T-1 adv-attack[13,16]	Trojaning attack[18-20]
投毒攻击	本文方法	Clean-label attack[10-12]

下面用一个简单的二分类问题说明两类投毒攻击的差异。此二分类器的任务是区分图像数字“3”和“8”。投毒攻击者的任务是篡改少量训练数据，使训练好的神经网络将特定的某个样本“3”错误地判别为类别“8”。为了实现此目标，可以利用第一类投毒攻击将部分训练数据“3”大幅度地改为“$\tilde{8}$”。其在人类眼中为数字 8，且拥有正确的类别“8”标签，但第一类投毒攻击能使训练后的网络将特定样本“3”错判为“8”。与之相对，现有的第二类投毒攻击[11-13]则是将部分训练数据“3”改为“$\tilde{3}$”，以实现相同的目标。

图 1 为第一类投毒攻击原理。由图 1 可以看出，第一类投毒攻击大幅改变数字“3”的图像，生成许多图像“$\tilde{8}$”。这些图像“$\tilde{8}$”在特征空间中环绕着原始“3”的特征分布。由于图像“$\tilde{8}$”被标注为“8”，故神经网络对该类投毒图像进行学习后，认为“3”也属于类别“8”，产生了错误预测。

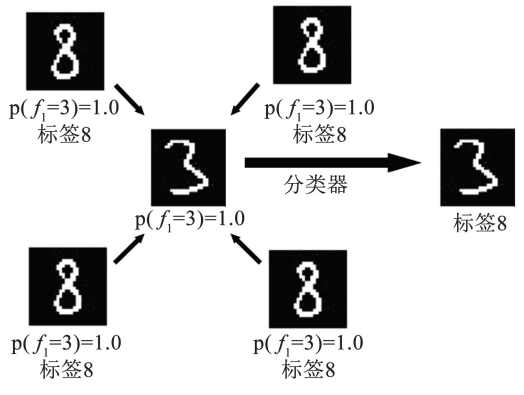

图 1　第一类投毒攻击原理

Fig. 1　Illustration of type I poisoning attack

为了更好地说明第一类投毒攻击与第二类投毒攻击对数据分布产生的影响，本文以数字“3”和“8”的分类任务为例，构建了一个数据空间分类示意图，见图 2，以辅助分析分类器在不同攻击策略下的决策行为。由图 2 可以看出，分类器最初正确地将数字“3”和“8”分开。但是，当数据集被注入投毒数据时，一些标签为数字“8”，但特征与数字“3”相似的样本出现在数字“3”的类别中（图 2 中用“o”表示的数据）。被这些投毒数据包围的数字“3”将被“拉走”（分类器认为其在特征空间中是“8”），并认为其应该被分类至数字“8”中。上述攻击方式即为第一类投毒攻击。

另外，还有一种投毒数据，其标签是数字“3”，但特征与数字“8”相似（图 2 中用“ * ”表示的数据）。这些数据包围的攻击目标数字“3”会因为附近的投毒数据而被分类器认为其在特征空间中具有相同的特征，从而认为攻击目标图像与数字“8”相似。此攻击方式即为第二类投毒攻击。这样，通过添加两种完全不同的投毒数据，完成了同样的投毒攻击任务。可以看到，当投毒数据被注入数据集后，数据集的分布发生了变化，导致分类器对原有数据集的分类边界发生了移动。分类边界发生变化后会影响被攻击目标的分类结果，最终导致投毒攻击生效。

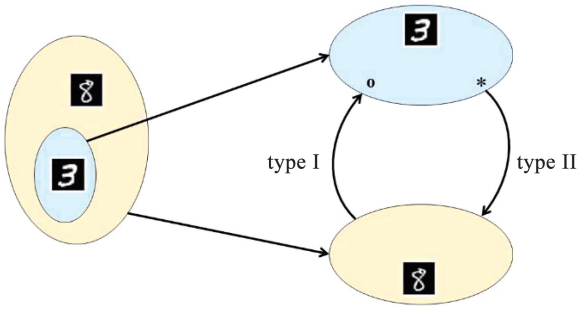

图 2　两类投毒攻击改变数据分布

Fig. 2　Data distribution changes under two types of poisoning attacks

与两类对抗攻击相似，两类投毒攻击也在理论上有着本质不同。第一类投毒攻击鼓励网络学习缺失的特征，而第二类投毒攻击鼓励网络学习冗余特征，这与第一类对抗攻击[13]中的论述是完全一致的。

2.2　问题描述

投毒攻击以隐蔽的手法污染训练数据，使训练好的模型在特定情况下判别错误。对于投毒攻击而言，攻击者需要制作投毒样本$\boldsymbol{x}_{\mathrm{p}}$，表达式为

$$\boldsymbol{x}_{\mathrm{p}}=\underset{\boldsymbol{x}}{\operatorname{argmin}}\|\boldsymbol{x}-\boldsymbol{x}_{\mathrm{b}}\|^{2}+\mu\|f(\boldsymbol{x})-f(\boldsymbol{x}_{\mathrm{t}})\|^{2} \tag{1}$$

式中：$\boldsymbol{x}$ 为原样本，$\boldsymbol{x}_{\mathrm{b}}$为基本类别，$\boldsymbol{x}_{\mathrm{t}}$为神经网络提取到特征的目标类别，$\mu$ 为权重参数，f 为分类器分类结果。

投毒样本在人眼看来更接近于基本类别$\boldsymbol{x}_{\mathrm{b}}$，然而，神经网络在训练过程中将其特征错误地关联至目标类别$\boldsymbol{x}_{\mathrm{t}}$，将其分类为目标类别，从而实现攻击效

果。这一过程会干扰神经网络对特征空间的学习，影响其分类边界的形成，从而导致模型对目标图像产生错误分类。

结合前述的二分类问题，第二类投毒攻击生成的样本在特征分布上与被攻击类别存在明显差异，但其外观通常被人类判定为属于目标类别。相比之下，第一类投毒攻击在生成投毒样本时选择的基本类别$\boldsymbol{x}_{\mathrm{b}}$与攻击目标类别不同，所构造的投毒样本在人类视觉判断下仍呈现为基础类别 $\boldsymbol{x}_{\mathrm{b}}$的典型特征。然而，受该攻击样本影响训练后的神经网络却会将其误判为目标类别，说明模型在学习过程中将基础类别与目标类别的特征错误地关联，从而实现攻击者设定的分类偏移。

在生成投毒样本的过程中，攻击者首先从测试集中选定一个目标样本作为本次攻击的具体对象。在一次成功的投毒攻击中，该目标样本将在测试阶段被模型错误分类。随后，攻击者从训练集中选择一个基本样本，并对其进行精细化修改，以生成投毒样本。该基本样本在视觉上与投毒样本高度相似，肉眼难以察觉其间差异，因此，在数据清洗或人工标注阶段容易被误认为标签正确。然而，从深度神经网络的角度看，攻击者通过优化算法对该图像的特征进行了隐蔽操控，使其在特征空间中逐渐逼近攻击目标样本的表示。最终，攻击者将该投毒样本注入模型的训练数据中，导致模型在学习过程中对类别边界产生错误更新，使得原本应归属于目标类别的测试样本被误判为基本样本所属类别。通过上述过程，攻击者成功诱导模型在不更改标签、不破坏样本外观的前提下，对指定目标样本做出错误预测，完成了一次具备高度隐蔽性与定向性的投毒攻击。

2.3 第一类投毒攻击

对于第一类投毒攻击样本，攻击者希望能够生成一个基于原样本 $\boldsymbol{x}$ 的投毒样本$\boldsymbol{x}'$。特别需要注意，攻击者希望被攻击分类器f_1（通常被认为是攻击者针对的深度神经网络）对这两个样本具有相同的识别结果，即深度神经网络对原样本和投毒样本在特征层面上都认为具有相同类别的特征。但对于专家分类器f_2（通常情况下是被专业训练过的人类）而言，却有着不同的分类结果。也就是说，攻击者希望生成的对抗样本对人类观察而言具有不同的形状。这个投毒样本会被人眼认为区别于原样本，从而被标记与其特征层面不同的分类标签。以上过程的表达式为

$$\begin{aligned} &\boldsymbol{x}' = A(\boldsymbol{x}) \qquad (2)\\ \text{s. t. } &f_1(\boldsymbol{x}') = f_1(\boldsymbol{x})\\ &f_2(\boldsymbol{x}') \neq f_2(\boldsymbol{x}) \end{aligned}$$

式中 $A(\boldsymbol{x})$为生成投毒样本编码器。式(2)展示了原样本 $\boldsymbol{x}$ 和投毒样本$\boldsymbol{x}'$的关系。

与第一类对抗攻击[13]相同，本文中的投毒样本$\boldsymbol{x}'$由监督变分自编码器自原样本 $\boldsymbol{x}$ 生成，如图 3 所示。其中，z、z_{p}分别为原样本、投毒样本的隐变量，J_{dis}为鉴别器的判断结果，J_{dec}为解码器的判断结果，J_{KL}为 KL 散度的判断结果，J_1、J_2分别为f_1、f_2的判断结果。特别地，本文利用 KL 散度测量原样本与投毒样本特征空间的距离，以保证两者具有相同的分类结果。最终，所生成的投毒样本在f_1和f_2上将产生不一致的分类结果，反映出两者在特征判别机制上的差异性。

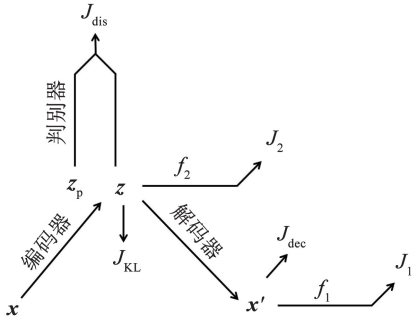

图 3 生成第一类投毒攻击样本框架

Fig. 3 Generate the type I poisoning attack samples

在生成投毒样本的过程中，f_1的梯度不仅像传统方法那样反向传播至f_1的输入$\boldsymbol{x}'$，而且通过解码器进一步反向传播至隐变量 z。同时，利用攻击者的梯度修正 z，通过解码器得到一个新的带有不同标签的图像$\boldsymbol{x}'$。在此过程中，f_1和f_2之间需要平衡，以保持f_1的输出不变，最终实现第一类攻击。

基于经过训练的监督变分自编码器，将从原始样本 $\boldsymbol{x}$ 生成一个第一类投毒样本$\boldsymbol{x}'$，使得$\boldsymbol{x}'$和 $\boldsymbol{x}$ 在f_2视图中具有不同的标签，但被f_1识别为同一类。该过程将输入图像 $\boldsymbol{x}$ 转换为另一个具有不同标签的图像$\boldsymbol{x}'$，实现图像转换任务。在该框架中，类别信息不是直接给出的，而是来自攻击者的监督项。隐变量根据f_2的梯度进行迭代修正，并通过解码器恢复成图像。从生成过程可以看出，第一类投毒攻击通过大幅度篡改少量训练数据，能够使神经网络在正确标签的指引下，无法完整学习正确的图像分布，使得其对特定图像产生极高的错误率。

通过以上方式，监督变分自编码器成功地生成了带有所需标签的新样本。此外，结合f_1，监督变分自编码器尝试保持f_1的输出不变，即生成第一类投毒样本。对于f_1，通过最小化以下函数生成具有原始标签 y 的输入 $\boldsymbol{x}$ 的目标标签为y'的投毒样本$\boldsymbol{x}'$。生成目标样本公式为

$$J_{SA}=J_{IT}+k_t J_1(\boldsymbol{x}',y,\cdot)=-y'\log f_2(z)+\alpha(1-f_{dis}(z))+k_t J_1(f_{dec}(z),y,\cdot)+\gamma\|z\|_2 \tag{3}$$

式中：J_{SA}为监督变分自编码器的损失函数，J_{IT}为图像转换任务的损失函数，k_t为压力参数，$J_1(\boldsymbol{x}'\cdot)$为被攻击分类器的损失函数，$\alpha$、$\gamma$为权重系数，$f_{dis}$为鉴别器函数，$f_{dec}$为解码器函数。在式(3)中，正参数$k_t$反映了为保持$f_1$不变而设置的损失函数强度。在生成第一类投毒样本时，k_t可以根据不同的迭代需求而变化。通常情况下，监督变分自编码器允许通过引导f_1的输出变化来驱动图像特征的逐步重构与偏移。随着训练过程的推进，k_t的值可以逐步增大，以加强对图像生成过程中类别一致性的约束，从而使投毒样本在f_1的判别视角下仍保持与原始基本样本相同的分类结果，确保攻击的目标性与隐蔽性。

2.4　基于特征空间的两类投毒攻击本质解析

人类与分类器在判别标准上的差异，是神经网络产生错误分类判断的根本原因之一。图 4 为两类投毒攻击区别的解释性模型，可用两个不同的超平面分别表示人类与分类器的决策边界。在这个解释性模型中，假设输入空间包含 3 个正交的特征方向：$x(1)$、$x(2)$和$x(3)$。其中，分类器主要基于$x(1)$和$x(3)$特征进行正负样本划分，而人类则依赖于$x(1)$和$x(2)$进行判断。对于人类，$x(3)$方向的特征为不必要的特征，因为人类并不以其作为判定标准。当样本在其分类边界上沿着$x(3)$方向移动时，人类是不能观察到新图像与原图像的区别的。因此，攻击者可以利用该特性，从负样本出发，在$x(3)$方向生成扰动样本，即投毒样本，并将其分布在目标样本的特征空间周围。对于分类器而言，这些精心构造的投毒样本会对原有的决策边界产生干扰，从而导致分类器在测试阶段将目标样本误判为负类。其利用了神经网络在不必要特征上的过度学习，分别在测试阶段和训练阶段完成了对神经网络的愚弄。

相比之下，$x(2)$方向的特征为缺失的特征，因为人类以其作为判别准则，而神经网络并未学会此特征。同样，可以沿着$x(2)$负方向制作投毒样本。这些投毒样本会被人类看作其他类别的图像，也就是越过了人类对图像的分类边界。但是对于分类器而言，投毒样本与原样本在$x(1)$和$x(3)$上没有发生变化，因此会被分类于原有特征空间中。这些样本就是第一类投毒样本。之后，这些投毒样本会被标签为其他类图案，包裹住目标样本，导致分类器改变原有的分类边界。最终，目标样本会被认为与投毒样本具有相同的标签。投毒样本利用神经网络对某些特征的忽视，成功愚弄了人类和分类器，实现了第一类投毒攻击。需要特别注意的是，这是一个理想的模型，在实际应用中，各个特征之间不是完全正交的，该解释性模型是为了更加直观展示第一类投毒攻击与第二类投毒攻击在理论实现方面具有不同的特性而进行的假设。

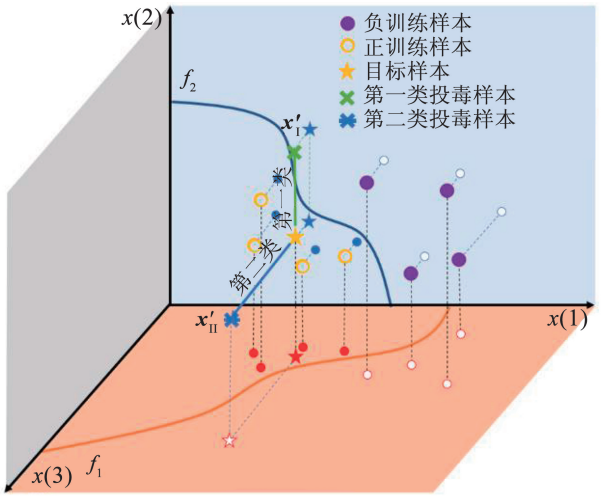

图 4　两类投毒攻击区别的解释性模型

Fig. 4　Explanatory model for differences between two poisoning attacks

第二类投毒攻击的本质在于，攻击者在投毒样本中有意引入部分冗余特征，使得神经网络在训练过程中错误地学习了这些非判别性特征。在测试阶段，攻击者可通过构造包含相同冗余特征的输入样本，诱导模型将本应正确分类的样本误判为其他类别。换言之，该攻击方式利用了模型对不必要特征的过度学习，使得模型在面对人类看来明显应归于某一类别的样本时，做出错误的分类决策。

在实际训练过程中，攻击者通过构造一组围绕目标样本分布的投毒样本，引导模型学习到错误的类别边界，使得原本仅属于投毒样本的特征也被认为是目标类别的有效特征。这种特征层面的包围改变了模型的决策边界，从而实现了定向干扰。相反，第一类投毒攻击的产生则源于模型在训练过程中未能充分学习关键判别性特征。当攻击者使用与目标类别明显不同的基本样本生成投毒图像时，由于图像在外观上与原样本差异较大，人工标注时会被赋予基本类别的标签。然而，神经网络可能因训练数据中缺乏足够的特征表达，无法准确区分二者，从而将目标样本误判为投毒样本所属类别。这一过程反映出神经网络对关键特征缺失所引发的泛化能力的不足。

本文提出的解释性模型，揭示了第一类投毒攻击与第二类投毒攻击在数据空间分布与特征操控机

制上的本质差异。基于两者与对抗攻击之间的耦合关系,本文认为第一类投毒攻击的提出为拓展现有投毒攻击研究范式提供了新路径,并可借助已有的第一类对抗攻击方法进行模型迁移与改进,从而增强攻击的理论完备性与实践灵活性。

从更高层次看,通过构造第一类对抗样本与第二类对抗样本,有望建立对抗攻击与投毒攻击之间的映射关系,进而将部分对抗攻击策略迁移应用于投毒攻击场景。基于统计学错误分类的思路可拓展出更多适应不同模型架构与任务环境的投毒攻击方式。现有第二类投毒攻击多集中于对某一特定目标样本的攻击设计,其攻击目标较为集中,易在实际系统中被针对性检测与防御。例如,在高安全等级的人脸识别系统中,关键目标人物类别的图像通常会被重点审查,此类攻击容易被人为清洗机制识别。相比之下,第一类投毒攻击通过污染数据库中其他类别图像,使其在模型学习中被错误地关联至目标类别,从而有效隐蔽攻击意图,提升了攻击的隐蔽性与通用性。第一类投毒攻击的策略在实际场景中更具操作空间,为未来深度神经网络系统的安全评估与防御机制设计提出了更具挑战性的问题。

3 实验与分析

实验主要验证两个问题:1)第一类投毒攻击的有效性;2)第一类投毒攻击与第二类投毒攻击是否有本质区别。

3.1 数据集

本文在 CIFAR-10[21] 数据集上验证第一类投毒攻击对多种常用神经网络的攻击成功率。只有当特定图像$\boldsymbol{x}_b$被归类为投毒目标样本类$\boldsymbol{x}_t$时,攻击才是成功的。对于所有实验,投毒样本生成时,按照4:1:1的比例将数据集划分为训练集、测试集和验证集。对网络投毒时,训练集的投毒样本比例为0.1%。

3.2 实验环境及参数设置

所有实验均在12 GB显存的NVIDIA TITAN X GPU上实现。对于每一个被攻击的神经网络,均进行577次攻击实验,并计算实验中的攻击成功率。在监督变分自编码器的训练中,采用Adam优化器[22]迭代优化隐变量$\boldsymbol{z}$,学习率为2×10^{-4}。α、γ分别设置为1×10^{-2}、1×10^{-4},与以往研究中的参数保持一致[13]。在攻击迭代过程中,利用Adam优化器更新隐变量$\boldsymbol{z}$,学习率为5×10^{-3}。

3.3 不同网络模型下第一类投毒攻击的有效性验证

为全面评估第一类投毒攻击方法在不同网络中的有效性,本文设计了一组针对主流深度神经网络模型的攻击实验。不同的神经网络在特征提取能力、结构深度、参数复杂度等方面存在差异,可能对投毒样本的学习与响应机制产生影响。因此,有必要在多个代表性网络上验证所提方法的有效性,以确保其具有良好的通用性和鲁棒性。

实验选取ResNet50[23]、ResNet18[23]、VGG16[24]、ConvNet[25]以及MobileNetV2[26]5种深度神经网络结构作为被攻击对象。这些网络在图像分类领域被广泛应用,涵盖了从轻量级网络到中大型残差网络的不同层级,能够较为全面地反映投毒攻击在不同结构条件下的表现差异。

实验结果见表2。可知,所有被测试的神经网络模型在无攻击条件下的训练准确率均超过92.00%,说明这些网络在正常数据分布下具有较强的学习能力和良好的分类性能。与此同时,本文所提出的第一类投毒攻击方法在这些主流深度神经网络模型上均展现出显著的攻击效果,攻击成功率普遍超过70.00%。说明即使在结构设计差异较大、参数量和容量不同的网络上,该方法仍具备较强的跨网络攻击能力。

表2 不同神经网络模型下第一类投毒攻击的攻击效果

Tab.2 Evaluation of type I poisoning attacks effectiveness across different neural network models

受害网络	准确率/%	攻击成功率/%
ResNet50	95.87	76.64
ResNet18	96.32	75.21
VGG16	94.90	71.64
ConvNet	93.71	79.57
MobileNetV2	92.58	77.46

这一结果充分验证了第一类投毒攻击方法的有效性和通用性,说明该方法能够在保持投毒样本标签一致性的前提下,显著干扰神经网络的判别边界,对训练充分、准确率较高的网络构成了实质性的安全威胁。此外,实验还表明,第一类投毒攻击不仅具备隐蔽性强、通过传统手段难以识别的特点,而且攻击稳定性较高。

3.4 基于第二类投毒防御方法的第一类投毒攻击鲁棒性分析

为进一步验证第一类投毒攻击与第二类投毒攻击在攻击机制上的本质差异,本文设计实验评估当前主流第二类投毒攻击防御方法在应对第一类投毒攻击时的有效性。当前已有的第二类投毒防御策略主要集中于数据增强方法,通过扰动图像结构,削弱

投毒样本中的微小扰动,从而提升模型鲁棒性。

实验选取高斯平滑滤波[27]、混合增强[28]、CutMix[29]和 CutOut[30]4 种已被证实在第二类投毒攻击下具有良好防御效果的方法。所有实验均在相同设置下进行,以 ResNet50 作为被攻击模型,对本文提出的第一类投毒攻击方法进行防御测试,结果见表 3。

表 3　基于现有防御方法的第一类投毒攻击有效性评估

Tab. 3　Effectiveness evaluation of type I poisoning attacks against existing defense methods

防御方法	防御成功率/%
高斯平滑滤波	17.85
混合增强	22.86
CutMix	19.53
CutOut	24.21

从表 3 可以看出,所选取的 4 种主流第二类投毒攻击防御策略在应对第一类投毒攻击时均未表现出理想的防御效果,防御成功率普遍偏低,最高值仅为 24.21%。这一结果表明,当前以数据增强为核心的第二类投毒防御机制难以在特征分布和攻击策略均存在显著差异的第一类投毒场景中发挥作用,说明其防御能力缺乏普适性和跨范式适应性。

综上所述,第一类投毒攻击与第二类投毒攻击在攻击机理、特征操控方式以及模型决策路径上的差异是根本性的。第二类攻击通常依赖于在保持图像语义一致的前提下微调特征以误导模型,而第一类攻击则通过大幅度修改样本结构,引导模型学习错误的特征边界。由于攻击原理存在本质不同,现有面向第二类投毒攻击的防御方法在面对第一类攻击时表现出明显的适应性缺失。这一发现表明,无法简单依赖现有防御体系覆盖所有类型的投毒攻击场景。

4　结　论

本文提出了第一类投毒攻击的定义,分析了其与第二类投毒攻击、第一类对抗攻击和第二类对抗攻击的本质区别,并基于监督变分自编码器生成了第一类投毒样本,通过实验验证了第一类投毒攻击能够有效干扰目标神经网络的学习过程,从而实现对模型的成功毒化。主要结论如下:

1)提出了第一类投毒攻击的概念,并从攻击目标与机制入手进行理论界定,指出其核心特征是在不更改数据标签的情况下,通过操控特征表示诱导模型产生第一类分类错误(即将正类误判为负类)。

2)提出了一种基于监督变分自编码器的第一类投毒样本生成方法,通过引导潜在空间扰动,在保持标签一致性的前提下诱导模型学习错误判别边界,从而在测试阶段引发第一类分类错误。

3)在多个主流神经网络上的实验表明,第一类投毒攻击在保持标签一致性的同时,可稳定诱导模型在测试阶段产生第一类分类错误,具有显著攻击效果与较强隐蔽性。

4)针对主流防御机制进行对抗性评估后发现,其在面对第一类投毒攻击时普遍失效,暴露出现有技术在特征鲁棒性建模上的局限,凸显该类攻击对模型安全体系的新挑战。

参考文献

[1]郭继昌, 郭昊, 郭春乐. 多尺度卷积神经网络的单幅图像去雨方法[J]. 哈尔滨工业大学学报, 2018, 50(3): 185
GUO Jichang, GUO Hao, GUO Chunle. Single image rain removal based on multi-scale convolutional neural network[J]. Journal of Harbin Institute of Technology, 2018, 50(3): 185. DOI: 10.11918/j.issn.0367-6234.201704075

[2]CHEN Qian, ZHU Xiaodan, LING Zhenhua, et al. Enhanced LSTM for natural language inference[C]//55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017: 1657. DOI:10.18653/v1/P17-1152

[3]COVINGTON P, ADAMS J, SARGIN E. Deep neural networks for YouTube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems. Boston: ACM, 2016: 191. DOI:10.1145/2959100.2959190

[4]CHENG H, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston: ACM, 2016: 7. DOI:10.1145/2988450.2988454

[5]SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations. Banff: ICLR, 2014

[6]SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-Gan: protecting classifiers against adversarial attacks using generative models[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018

[7]AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410. DOI:10.1109/ACCESS.2018.2807385

[8]XIANG Zhen, MILLER D J, KESIDIS G. A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense[C]//IEEE 29th International Workshop on Machine Learning for Signal Processing. Pittsburgh: IEEE, 2019: 1. DOI:10.1109/MLSP.2019.8918908

[9]BIGGIO B, NELSON B, LAVEL P. Poisoning attacks against support vector machines[C]//Proceedings of the 29th International Conference on Machine Learning. Edinburgh: Omnipress, 2012: 1807

[10] LIU Yanpei, CHEN Xinyun, LIU Chang, et al. Delving into transferable adversarial examples and black-box attack [C]//5th International Conference on Learning Representations. Toulon: ICLR, 2017

[11] SHAFAHI A, HUANG R, NAJIBI M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks [C]//32nd Conference on Neural Information Processing Systems. Montreal: NIPS, 2018: 6106

[12] TURNER A, TSIPRAS D, MADRY A. Clean-label backdoor attacks [C]//7th International Conference on Learning Representations. New Orleans: ICLR, 2019

[13] TANG Sanli, HUANG Xiaolin, CHEN Mingjian, et al. Adversarial attack Type I: cheat classifiers by significant changes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 1100. DOI:10.1109/TPAMI.2019.2936378

[14] STUTZ D, HEIN M, SCHIELE B. Disentangling adversarial robustness and generalization[C]//32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 6976. DOI:10.1109/CVPR.2019.00714

[15] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015

[16] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017: 39. DOI: 10.1109/SP.2017.49

[17] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018

[18] LIU Yingqi, MA Shiqing, AAFER Y, et al. Trojaning attack on neural networks[C]//25th Annual Network and Distributed System Security Symposium. San Diego: The Internet Society, 2018: 1. DOI: 10.14722/ndss.2018.23291

[19] JI Yujie, ZHANG Xinyang, JI Shouling, et al. Model-reuse attacks on deep learning systems[C]//25th ACM Conference on Computer and Communications Security. Toronto: ACM, 2018: 349. DOI: 10.1145/3243734.3243757

[20] REN Kui, MENG Quanrun, YAN Shoukun, et al. Survey of artificial intelligence data security and privacy protection [J]. Chinese Journal of Network and Information Security, 2021, 7(1): 1. DOI: 10.11959/j.issn.2096-109x.2021001

[21] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009

[22] KINGMA D P, BA J L. Adam: a method for stochastic optimization [C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015

[23] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770. DOI:10.1109/CVPR.2016.90

[24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015

[25] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(4): 541. DOI:10.1162/neco.1989.1.4.541

[26] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]//31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 4510. DOI:10.1109/CVPR.2018.0474

[27] BORGNIA E, GEIPING J, CHEREPANOVA V, et al. DP-InstaHide: provably defusing poisoning and backdoor attacks with differentially private data augmentation [C]//10th International Conference on Learning Representations. Appleton: ICLR, 2022

[28] ZHANG Hongyi, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018

[29] YUN S, HAN D, OH S J, et al. Cutmix: regularization strategy to train strong classifiers with localizable features[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 6023. DOI:10.1109/ICCV.2019.00612

[30] CUBUK E D, ZOPH B, MANE D, et al. AutoAugment: learning augmentation strategies from data [C]//32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach: IEEE, 2019: 113. DOI: 10.1109/CVPR.2019.00020