计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330521
61(5): 1182−1192, 2024

基于感知相似性的多目标优化隐蔽图像后门攻击

朱素霞　王金印　孙广路
（哈尔滨理工大学计算机科学与技术学院　哈尔滨　150080）
（黑龙江省智能信息处理及应用重点实验室（哈尔滨理工大学）　哈尔滨　150080）
（zhusuxia@hrbust.edu.cn）

Perceptual Similarity-Based Multi-Objective Optimization for Stealthy Image Backdoor Attack

Zhu Suxia, Wang Jinyin, and Sun Guanglu
（*School of Computer Science and Technology*, *Harbin University of Science and Technology*, *Harbin* 150080）
（*Heilongjiang Key Laboratory of Intelligent Information Processing and Application* (*Harbin University of Science and Technology*), *Harbin* 150080）

Abstract　Deep learning models are vulnerable to backdoor attacks and behave normally when processing clean data, but they will exhibit malicious behavior when processing toxic samples with trigger patterns. However, most backdoor attacks currently produce backdoor images that are easily perceived by the human eye, resulting in insufficient stealthiness of backdoor attacks. Therefore, a multi-objective optimized covert image backdoor attack method based on perceptual similarity is proposed. Firstly, the visual difference between the backdoor image and the original image is reduced using a perceptual similarity loss function. Secondly, a multi-objective optimization method is used to solve the problem of inter-task conflict on the poisoning model, thus ensuring stable performance of the model after poisoning. Finally, a two-stage training method is adopted to automate the generation of trigger patterns and improve the training efficiency. The final experimental results show that it is difficult for human eye to distinguish the generated backdoor image from the original image without any degradation in clean accuracy. Meanwhile, the backdoor attack is successfully performed on the target classification model, and the attack success rate reaches 100% for all experimental datasets under the all-to-one attack strategy. Compared with other steganographic backdoor attack methods, our method has better stealthiness.

Key words　backdoor attack; covert backdoor; poisoning attack; deep learning; model security

摘　要　深度学习模型容易受到后门攻击，在处理干净数据时表现正常，但在处理具有触发模式的有毒样本时会表现出恶意行为. 然而，目前大多数后门攻击产生的后门图像容易被人眼察觉，导致后门攻击隐蔽性不足. 因此提出了一种基于感知相似性的多目标优化隐蔽图像后门攻击方法. 首先，使用感知相似性损失函数减少后门图像与原始图像之间的视觉差异. 其次，采用多目标优化方法解决中毒模型上任务间冲突的问题，从而确保模型投毒后性能稳定. 最后，采取了两阶段训练方法，使触发模式的生成自动化，提高训练效率. 最终实验结果表明，在干净准确率不下降的情况下，人眼很难将生成的后门图像与原始图像区分开. 同时，在目标分类模型上成功进行了后门攻击，all-to-one 攻击策略下所有实验数据集的攻击成功率均达到了 100%. 相比其他隐蔽图像后门攻击方法，具有更好的隐蔽性.

收稿日期：2023-06-19；**修回日期**：2024-01-25
基金项目：黑龙江省自然科学基金项目 (LH2021F032)；黑龙江省重点研发计划项目 (2022ZX01A34)
This work was supported by the Natural Science Foundation of Heilongjiang Province (LH2021F032), and the Key Research and Development program of Heilongjiang Province (2022ZX01A34).
通信作者：孙广路（sunguanglu@hrbust.edu.cn）

关键词 后门攻击；隐蔽后门；投毒攻击；深度学习；模型安全

中图法分类号 TP391

近年来，深度神经网络得到了迅猛的发展和广泛应用，在人脸识别、自动驾驶、机器翻译、自然语言处理等众多领域表现出了优异的应用前景. 然而，训练这些网络需要大量的数据和计算资源，例如最近在多项自然语言处理任务上表现出色的 GPT-3 模型[1] 使用高达 1 750 亿个参数. 因此，许多用户选择将整个训练过程外包给云服务平台或采用迁移学习的方式进行训练. 然而，这些外包场景也降低了用户在模型训练过程中的控制权，攻击者可以轻松地对模型进行修改或破坏[2]，导致深度学习模型在安全性方面存在潜在的巨大风险.

后门攻击是一类通过毒害模型使其具有隐蔽后门功能的技术，目标是迫使中毒模型将特定的触发模式与攻击者指定的输出相绑定[3-5]. 攻击者通过向模型训练数据集中添加有毒样本，经过训练的模型在处理不包含触发模式的输入时表现正常；处理带有触发模式的输入时会被误导并执行错误的输出. 文献 [2] 提出的 BadNets 是深度学习神经网络中成功实现后门攻击的开创性研究，为后门攻击领域的发展奠定了重要基础，其描述的后门攻击步骤与方式被很多后续工作借鉴并进一步改进.

随着后门攻击的不断发展，攻击能力和强度也在不断增加. 简单后门攻击使用明显的后门模式已经不能满足实际需求，许多研究者借鉴了对抗样本[6]、信息隐藏[7] 和图形处理[8] 等领域的技术研究更加隐蔽的后门攻击. 隐蔽后门攻击旨在通过隐藏触发模式来增强隐蔽性，生成的后门图像难以被人眼察觉[9].

后门攻击图像隐蔽性与对抗性示例的研究方向[10] 类似，旨在通过增强后门图像与原始图像之间的相似性，从而提高攻击和抗检测能力. 文献 [11] 探索了隐蔽的后门攻击方法，为了确保攻击的隐蔽性，仅使用一定比例的触发模式叠加在干净图像上. 文献 [7] 提出了 2 种方法：第 1 种利用图像隐写技术将触发模式嵌入到比特位空间；第 2 种将 L_p 正则化约束得到的扰动增量作为触发器. 文献 [9] 在生成后门样本时考虑在像素空间中接近原样本，在特征空间中接近添加触发模式的样本. 然而，在增强隐蔽性的同时，后门攻击的准确率却开始下降. 例如，文献 [7] 的实验在 MNIST 数据集上对于数字“3”的攻击准确率小于 75%，对于数字“2”的攻击准确率小于 70%. 文献 [12] 提出使用一个小而平滑的扭曲形变来生成后门图像，因为人眼不善于识别较小的几何变换，这样可以使得对原始图像的改动不被注意到. 另外，文献 [13] 提出使用噪声作为触发模式的后门攻击方法.

尽管以上方法都在后门图像隐蔽性方向进行了相关工作，但由于在设计触发模式时缺乏约束后门图像与原始图像的相关指标，后门图像与原始图像之间仍存在着明显的差异，这种差异很容易被人眼感知发现，从而判断模型是否遭受到了攻击. 图 1 展示了 2 项最新工作，WaNet[12] 与 LIRA[13] 方法生成的后门图像与原始图像的对比可以看出，WaNet 生成的交通标志后门图像边缘形变比较严重，而 LIRA 方法触发模式生成器生成的噪声十分明显，导致后门图像带有明显的噪声，人眼很容易识别出来.

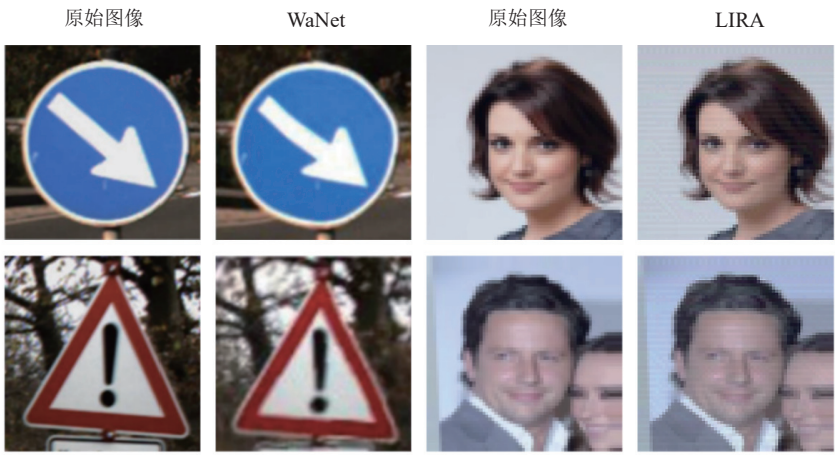

Fig. 1 Visualization of original images and generated backdoor images in WaNet and LIRA

图 1 WaNet 与 LIRA 中原始图像与产生的后门图像的可视化

为了解决深度学习网络图像分类任务中后门攻击图像隐蔽性不足的问题，本文提出了一种基于感知相似性的多目标优化隐蔽图像后门攻击方法PMOA，该法可以自主学习触发模式生成器，向被毒害分类模型植入一个不影响性能的后门，同时生成的后门图像与原始图像之间的区别无法被人眼察觉. 受到感知相似性的启发，本文首先基于感知距离计算后门图像与对应原始图像间的相似性，约束后门图像生成器的学习，解决了先前后门图像隐蔽性不足的问题，使人眼感知难以区分生成的后门图像与对应的原始图像；其次，通过多目标优化解决了中毒分类器模型在干净子集任务与有毒子集任务之间存在冲突的问题；最后，采用了两阶段训练方案，使得后门图像生成器的学习变得自动、高效且难以检测. 实验结果表明，本文方法在不降低被毒害模型原本分类准确率的前提下，显著提升了后门攻击的隐蔽性，对于开展后门攻击安全防御工作具有重要的参考价值. 本文的主要贡献包括3个方面：

1)为了使生成的后门图像具有很好的保真度，难以被人眼视觉检测出来，本文提出了一种基于感知相似性的后门图像生成器训练方法，该方法结合感知相似性损失计算，在不影响后门攻击准确率的前提下，生成的后门图像具有很高的保真度，提升了后门攻击的图像隐蔽性.

2)为了解决训练过程中被毒害分类器可能会陷入局部最小值而无法达到最佳性能的问题，本文采取了一种多任务目标优化方法，解决了正常任务学习和恶意任务学习之间产生冲突的问题，使得被毒害的分类器模型性能得到提高，从而提高后门攻击的鲁棒性.

3)使用多种不同的数据集和攻击设置进行了实验验证，体现了本文方法具有良好的性能，并且能够在各种环境下达到预先设定的攻击效果.

1 背　　景

1.1 基本的后门攻击方法

本文主要探讨有监督图像识别分类任务中的后门攻击方法. 标准的监督学习任务中，每个实例都由一个输入对象和一个期望的输出值组成，学习算法会分析训练数据并产生一个推断的功能，将新的实例映射到一个输出值，即学习出以ω为参数的映射函数$f_\omega: X \to Y$，其中$X=\{x_1,x_2,\cdots,x_N\}$表示输入域含有N个图像，$Y=\{y_1,y_2,\cdots,y_K\}$表示K个目标类的集合.

在正常情况下，训练图像分类任务的用户希望在训练集$D=\{(x_i,y_i): x_i\in X, y_i\in Y, i=1,2,\cdots,N\}$上训练出一个以$\omega$为参数的分类模型$f_\omega$，使得对于每张输入图像$x_i\in X$，都能够正确预测它所对应的标签$y_i\in Y$. 然而，攻击者可能会试图生成模型的攻击性训练集$D_{\text{attack}}$，从而导致训练出的模型$f_{\omega^*}$产生错误分类. 具体来说，攻击者需要2个子集：1) 干净子集$D_{\text{c}}=(X_{\text{c}},Y_{\text{c}})$，其中所有的样本都是正确的；2) 有毒子集$D_{\text{p}}=(X_{\text{p}},Y_{\text{p}})$，其中图像与标签都被攻击者改变. 通过将这2个子集合并起来，攻击者可以构造出攻击性训练集D_{attack}.

为了生成有毒的后门样本子集D_{p}，攻击者需要将原始训练样本(x_i,y_i)转换成带有后门的恶意样本$(T_\sigma(x_i),\eta(y_i))$. 其中$T_\sigma$表示以$\sigma$为参数的后门图像生成器，可以将原始图像$x_i$转换为带有后门的恶意图像$T_\sigma(x_i)$；$\eta$表示目标标签函数，可以将原始的标签$y_i$转换为攻击者指定的标签$\eta(y_i)$. 当攻击者使用调配好的$D_{\text{attack}}$来训练分类模型$f_\omega$时，可以改变模型的行为，并使被毒害的分类器$f_{\omega^*}$将带有特殊触发模式的输入数据分类到攻击者指定的目标类别$\eta(y_i)$，而不带有触发模式的干净输入数据则会正常分类到y_i. 具体为：

$$f_{\omega^*}(x_i)=y_i, \tag{1}$$

$$f_{\omega^*}\left(T_{\sigma^*(\omega)}(x_i)\right)=\eta(y_i). \tag{2}$$

通常，后门攻击设置根据目标标签函数η的不同可分为2种：多对1攻击(all-to-one)和多对多攻击(all-to-all). 在all-to-one攻击中，攻击者将标签设置为一个固定的常量目标，即所有后门输入样本的标签是一致的：$\eta(y_i)=c, i=1,2,\cdots,N$，这种攻击方式也被称为单目标攻击. 对于all-to-all攻击，攻击者会将标签设置为相对于真实标签的一个偏移量，即每一类中毒图像均对应不同的目标标签：$\eta(y_i)=(y_i+1)\bmod|c|$，$i=1,2,\cdots,N$，其中有毒的子集样本每类的标签都被更改为总共$K$个类别中的另一类，且转化后的标签类之间不重合.

后门图像生成器T_σ将原始的输入x映射到一个新的后门输入$T_\sigma(x)$，通过带有后门功能的模型处理后门输入的结果为$\eta(y)$. 为了让被毒害的分类模型更好地区分原始图像与后门图像，先前的后门攻击通常采用异常明显的后门模式，导致原始图像与后门图像之间差异过大，人眼可以轻松地识别有毒数据. 因此，研究者们开始关注后门攻击的隐蔽性，即如何实现不易被人眼发现的后门攻击.

隐蔽后门攻击通过提高后门图像的视觉质量从而增强对人类检查的隐蔽性. 这种隐蔽攻击的后门

图像生成器定义为：

$$T_\sigma(x)=x+g_\sigma(x). \tag{3}$$

目标是训练出一个以σ为参数、人眼无法感知的扰动生成模型g_σ，旨在最小化被毒害的分类模型在x和$T_\sigma(x)$上的损失函数. 换言之，在给定α和β这2种任务权重参数的情况下，需要最小化损失函数：

$$\min_{\substack{\theta\\ \alpha\in[0,1]}}\sum_{i=1}^{N}\alpha L(f_{\omega^*}(x_i),y_i)+\beta L\left(f_{\omega^*}\left(T_{\sigma^*(\omega)}(x_i)\right),\eta(y_i)\right). \tag{4}$$

1.2 威胁模型

本文考虑实际场景中最常见的攻击方式，假设后门功能是在分类模型训练时注入的. 攻击者可以在模型训练前完全访问被毒害的模型，包括模型结构和参数，但是在训练进行中，攻击者没有关于训练细节的信息，也无权更改训练组件，包括训练损失、模型结构和训练计划. 在测试阶段，攻击者可以使用后门图像测试被毒害的模型，但是无法控制被毒害模型的推理过程，也无法访问被毒害的模型.

攻击者的主要目标是后门攻击的隐蔽性和有效性. 首先需要保证生成的后门图像是从干净的图像中制作出来的，没有明显的修改，无法被人眼察觉到. 其次，被毒害的分类器f不应该降低原本在干净子集上的分类性能，同时保证有毒子集上的性能最大化. 具体总结为2点：

1）干净的原始图像与它的后门图像仅仅存在着人眼无法感知的细微差别；

2）被毒害的分类器与它的原始版本表现一致，将干净数据x分类到正常标签y，但对其后门图像$T(x)$的预测会更改为$\eta(y)$.

2 基于感知相似性的后门学习方法

先前的隐蔽后门攻击在生成后门图像时缺乏与原始图像的约束指标，产生的后门图像效果未能达到预期，经过对比后很容易被人眼感知到. 因此，本文基于感知相似性提出了一种可以避免人眼感知察觉的后门触发模式，训练出了一个优秀的后门图像生成器T. 同时为了避免干净子集的分类任务与有毒子集的后门任务之间的冲突，基于多目标优化提出了一种新的模型后门注入训练方法.

2.1 威胁模型感知相似性计算

目前的后门攻击模型缺乏有效的约束性指标来确保生成的后门图像既能够激活后门功能，又具有良好的保真度，不会被人眼轻易地与原始图像区分开. 经典的图像质量评估方法通常假设像素间相互独立来计算损失，不能考虑到像素之间的上下文信息，无法捕捉到2张图像之间的感知差异. 为了避免被人类视觉检查发现，隐蔽后门攻击的后门图像生成器需要一种符合人类判断方式的相似度衡量方法，以更好地判断后门图像是否会被发现.

最新研究表明[14-16]，许多图像转换任务可以使用感知损失函数生成高质量的图像，不再采取计算独立像素之间差异的方法，而是从卷积神经网络中提取图像特征表示像素之间的差异来进行衡量，并通过最小化损失函数生成图像. 采用深度神经网络实现该目标也非常有效，尽管网络的内部激活是针对图像分类任务进行训练的，但作为特征空间仍然具有极高的效率. 文献[17]通过从深度学习模型的网络结构中提取图像的高层特征，并计算特征之间的感知距离（perceptual distance）来度量感知相似性. 这种方法更符合人类感知的判断方式，能够更好地衡量图像之间的相似度，比基于逐个像素比较的图像评估方法更加出色.

为了获取2张图像的感知距离与相似性之间的关系，需要训练一个大规模的卷积神经网络，通过在大量图像数据集上进行无监督或有监督的训练，最终得到一个强大的特征提取器. 对于2张需要计算感知差异的图像，将它们输入到特征提取器中，得到它们在某些中间层的特征表示. 在特征空间中计算这2个特征表示之间的距离，使用网络G计算距离得分作为2张图像之间感知相似性的度量.

受到感知距离计算相似性的启发，针对图像分类任务进行后门攻击，本文提出应用感知相似性解决后门图像隐蔽性不足的问题. 在给定一个被毒害神经网络f的情况下，计算原始图像x与后门图像$T_\sigma(x)$之间的感知距离d，最后通过感知判断网络，计算二者之间的感知相似性作为约束训练后门图像生成器. 为了计算后门图像与原始图像之间的感知距离，首先需要计算深度嵌入，将通道维度上的激活归一化，使用向量$\boldsymbol{w}$缩放每个通道，计算欧式距离. 其次，在空间维度和所有层之间进行平均，获得原始图像x与后门图像$T_\sigma(x)$之间的感知距离d，如图2所示.

具体来说，本文计算后门图像与对应原始图像之间的感知距离需要从L层中提取特征堆栈，并在通道维度上进行单位归一化，对于L层，将得到的结果记为$\hat{\boldsymbol{y}}^l_{(x)},\hat{\boldsymbol{y}}^l_{(T_\sigma(x))}\in\mathbb{R}^{H_l\times W_l\times C_l}$. 利用向量$\boldsymbol{w}^l\in\mathbb{R}^{C_l}$缩放激活通道并计算欧式距离. 最后，通过在空间和信道上分别取平均值与求和获得距离：

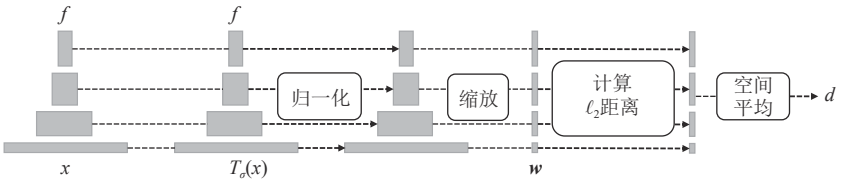

Fig. 2 Calculation of perceptual distance

图 2 感知距离的计算

$$d(x,T_\sigma(x))=\sum_l\frac{1}{H_lW_l}\sum_{h,w}\left(\boldsymbol{w}^l\odot\left(\hat{\boldsymbol{y}}^l_{(x)hw}-\hat{\boldsymbol{y}}^l_{(T_\sigma(x))hw}\right)\right)_2^2. \tag{5}$$

2.2 隐蔽的后门学习

在本文的隐蔽后门攻击任务中，后门图像生成器T_σ的学习非常重要，因为它必须保证在不影响干净图像分类准确率的情况下最大化后门攻击准确率，并且保持隐蔽以避免被人类防御者所发现. 感知相似性损失[17]是一种基于感知距离来模拟人眼检测，评估两幅图像相似度的度量指标，损失值越低，说明两幅图像之间的相似度越高. 为了训练后门图像生成器，本文将后门图像与原始图像的感知相似性损失和在被毒害分类器上的分类损失相结合，不断训练后门图像生成器，可以实现人眼无法感知的隐蔽后门攻击.

如图 3 所示，本文的隐蔽后门图像生成器学习系统由 3 部分组成：扰动生成网络g_σ、图像分类网络f_ω和感知判断网络G组成. 扰动生成网络是一个由σ参数化的自编码器，通过映射$T_\sigma(x_i)^{eps}=x_i+g_\sigma(x_i)\times eps$将输入图像与对输入图像的一部分扰动相结合作为训练后门准确率的后门图像$T_\sigma(x_i)^{eps}$，其中$g_\sigma(x_i)$表示对输入图像x_i生成的扰动，eps表示投毒率；同时通过映射$T_\sigma(x_i)=x_i+g_\sigma(x_i)$将输入图像与对输入图像的全部扰动相结合作为训练后门隐蔽性的后门图像$T_\sigma(x_i)$. 接下来从图像分类网络中提取原始图像x_i与后门图像$T_\sigma(x_i)$的感知距离，并通过感知判断网络计算二者之间的感知相似性$\ell_{\text{perceptual}}$. 最后结合后门图像$T_\sigma(x_i)^{eps}$在分类器上的后门损失ℓ_{poison}训练后门图像生成器T_σ.

为了确保后门攻击的隐蔽性，需要在输入空间上产生人眼不易察觉的扰动，这可以通过最小化后门图像与原始图像之间的感知相似性损失来实现. 此外，还需要最小化后门子集在分类器模型上的后门损失来最大化攻击性能. 即：

$$\min\sum_{i=1}^{N}\{-h_i\log G(d(x_i,T_\sigma(x_i)))-(1-h_i)\log(1-G(d(x_i,T_\sigma(x_i))))\}, \tag{6}$$

$$\min\sum_{i=1}^{N}L\left(f_{\omega^*}\left(T_{\sigma^*(\omega)}(x_i)\right),\eta(y_i)\right). \tag{7}$$

触发模式生成器可以被设计为简单的自编码器，也可以使用更复杂的 U-Net 架构[12]. 然而，本文进行的实验发现自编码器和 U-Net 之间的性能差异并不显著. 另外，考虑到 U-Net 结构更加复杂，需要更长的训练时间. 因此，本文在后续实验中选择使用自编码器作为扰动生成模型的实现方式.

2.3 多任务目标优化的分类器模型训练

在对分类器模型进行后门攻击时，由于后门图

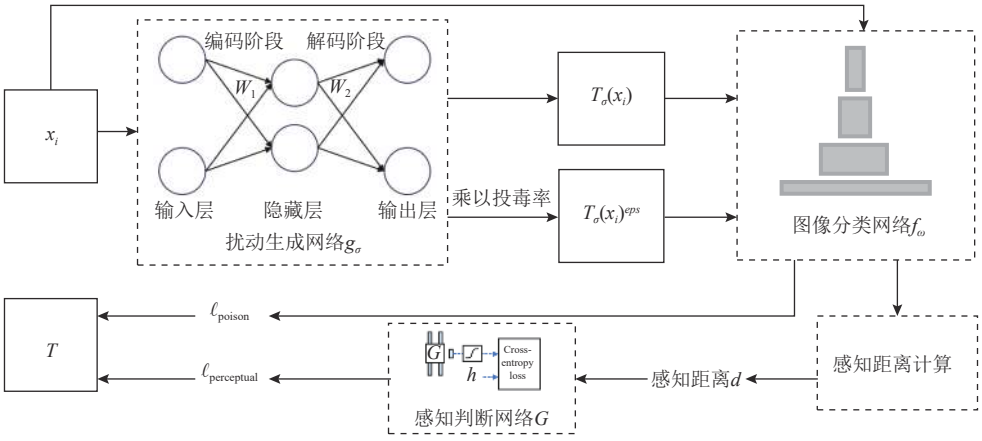

Fig. 3 Learning process of backdoor image generator

图 3 后门图像生成器学习过程

像与原始图像之间相似度较高，干净子集分类任务与有毒子集后门任务之间存在天然的冲突，可能无法同时得到充分优化，导致其中一个任务在训练过程中占主导地位，而另一个任务无法达到最佳性能. 结果是模型无法收敛到全局最优解，只能停留在局部最优解处.

为了避免这种情况出现，许多研究实验将式(4)中2种任务的损失强度α和β均设置为0.5[13]，以期望被毒害的分类器可以在2种任务上都达到最佳性能. 但是根据本文进行的实验观察到，受攻击的分类器在不同的数据集下的学习情况是不同的，通用的设置可能会导致深度神经网络无法达到预期的最佳状态或者性能不稳定. 因此，需要在这2种分类任务之间进行权衡，本文希望最小化2个任务在被毒害分类模型上的损失为：

$$\min_{\substack{\theta\\ \alpha\in[0,1]}} \sum_{i=1}^{N} \alpha L(f_{\omega^*}(x_i),y_i)+ (1-\alpha)L\left(f_{\omega^*}\left(T_{\sigma^*(\omega)}(x_i)\right),\eta(y_i)\right). \tag{8}$$

式(4)可以理解为期望在中毒模型上进行多任务学习，包括模型原本在干净数据集上的分类任务，以及攻击者计划在有毒数据集上的后门任务. 实际上，多任务学习的本质是一个多目标问题，因为干净子集分类任务与有毒子集后门任务之间存在冲突. 因此，在进行后门攻击时需要考虑如何对这2个任务之间进行优化.

通常情况下，多任务学习的优化办法是优化一个代理目标，以使得每个任务损失的加权线性组合最小化，但这种方法只在多个任务间不存在相互竞争的情况下才有效. 本文的干净子集分类任务与后门任务之间存在相互竞争，因此优化一个代理目标可能不是最佳的策略. 文献[18]中将多任务学习的优化当作多目标优化的问题来处理，寻找帕累托最优解(Pareto optimality)来优化多任务学习. 帕累托最优解是指在多个目标之间不存在可比性时所有最佳解的集合. 本文遵循这一思路，希望计算出2种任务之间的最优权重解α.

首先考虑一个在输入空间X和一系列任务空间$\{Y^t\}_{t\in[T]}$集合上的多任务学习，预测函数可以表示为$f^t(x;\theta^{\mathrm{sh}},\theta^t):X\to Y^t$，$\theta^{\mathrm{sh}}$表示不同任务之间共享的参数，$\theta^t$是任务相关的参数. 多任务学习求解的一般形式表示为：

$$\min_{\substack{\theta^{\mathrm{sh}}\\ \theta^1,\cdots,\theta^T}} \sum_{t=1}^{T} c^t \hat{L}^t(\theta^{\mathrm{sh}},\theta^t) \tag{9}$$

其中c^t是对于每个任务的静态或动态计算的权重，$\hat{L}^t(\theta^{\mathrm{sh}},\theta^t)$是任务$t$的经验损失，定义为$\hat{L}^t(\theta^{\mathrm{sh}},\theta^t)\triangleq\frac{1}{N}\sum_i L(f^t(x_i;\theta^{\mathrm{sh}},\theta^t),y_i^t)$. 使用矢量值的损失$L$来指定MTL的多目标优化表述，定义为

$$\min_{\substack{\theta^{\mathrm{sh}}\\ \theta^1,\cdots,\theta^T}} L(\theta^{\mathrm{sh}},\theta^1,\cdots,\theta^T)=\min_{\substack{\theta^{\mathrm{sh}}\\ \theta^1,\cdots,\theta^T}} (\hat{L}(\theta^{\mathrm{sh}},\theta^1),\cdots,\hat{L}(\theta^{\mathrm{sh}},\theta^T))^T. \tag{10}$$

文献[19]提出了一种多梯度下降算法，利用满足KKT(Karush Kuhn tucker)条件的帕累托静止点，将优化问题进行转换，然后求解以上优化问题. 将得到的解$\sum_{t=1}^{T}\alpha^t\nabla_{\theta^{\mathrm{sh}}}$作为梯度更新应用于共享参数.

对于干净子集分类任务与后门任务而言，上述的优化问题会转换为：

$$\min_{\alpha\in[0,1]} \alpha\nabla_{\omega^{*\mathrm{sh}}}L^1(\omega^{*\mathrm{sh}},\omega^{*1})+(1-\alpha)\nabla_{\omega^{*\mathrm{sh}}}L^2(\omega^{*\mathrm{sh}},\omega^{*2})_2^2, \tag{11}$$

其中L^1和L^2均为损失函数

这是一个具有分析解的α的一维二次函数，其解为：

$$\hat{\alpha}=\left[\frac{\left(\nabla_{\omega^{*\mathrm{sh}}}\hat{L}^2\left(\omega^{*\mathrm{sh}},\omega^{*2}\right)-\nabla_{\omega^{*\mathrm{sh}}}\hat{L}^1\left(\omega^{*\mathrm{sh}},\omega^{*1}\right)\right)^T\nabla_{\omega^{*\mathrm{sh}}}\hat{L}^2\left(\omega^{*\mathrm{sh}},\omega^{*2}\right)}{\nabla_{\omega^{*\mathrm{sh}}}\hat{L}^1\left(\omega^{*\mathrm{sh}},\omega^{*1}\right)-\nabla_{\omega^{*\mathrm{sh}}}\hat{L}^2\left(\omega^{*\mathrm{sh}},\omega^{*2}\right)_2^2}\right]_{+,\frac{1}{T}}, \tag{12}$$

其中$[\bullet]_{+,\frac{1}{T}}$代表对[0,1]剪接到$[a]_{+,\frac{1}{T}}=\max(\min(a,1),0)$.

训练被毒害分类器f_{ω^*}时可以将干净数据集任务和有毒数据集任务权重的求解转换为帕累托最优，并最终通过多任务目标优化计算出干净数据集任务损失权重α和有毒数据集任务的损失权重$1-\alpha$. 这样使得分类器不会在其中一个数据集上过早地收敛，保证了分类器在2个数据上的性能都会达到最优.

2.4 模型交替训练

对于后门图像生成器T的学习，如果像生成对抗网络(generative adversarial network, GAN)[20]的训练方式一样，在干净数据集和有毒数据集上同时训练分类器f，只在有毒数据集上训练后门图像生成器T，很可能会出现分类器与后门图像生成器难以同时收敛、收敛时间过长或训练过程陷入局部最小值等问题. 因为这种训练方式要求双方在博弈中达到平衡，但是其中一方损失的下降都可能会导致另一方损失的上升.

因此，本文将后门图像生成器T与分类器f的学习同步进行，采取模型交替训练方案设计了对于分类器模型的投毒与后门图像生成器的训练，如图4

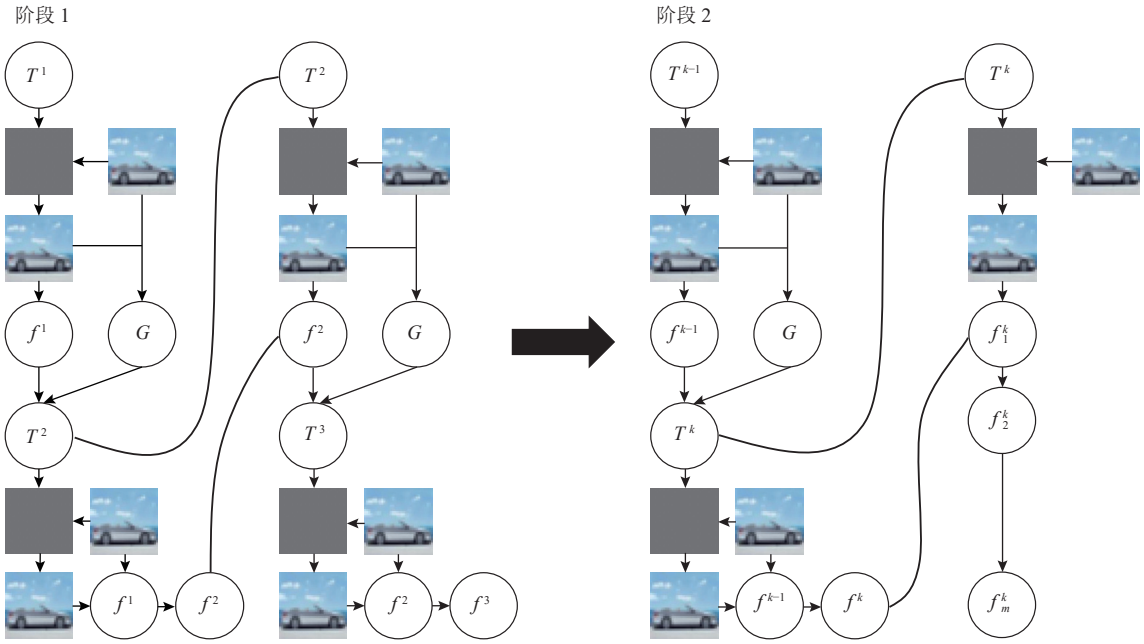

Fig. 4 Model training process

图 4 模型训练过程

所示. 具体来说，阶段 1 首先通过感知相似性损失与分类器对于有毒数据的分类损失训练后门图像生成器T. 其次使用训练后的后门图像生成器，采用多目标优化方法根据分类器f给出对干净数据与有毒数据的损失来训练分类器f. 如此循环 k 次，训练后门图像生成器 T 并毒害分类器f. 阶段 2 使用阶段 1 训练完毕的后门图像生成器T^k对分类器f^k进行微调.

同时学习分类器f和后门图像生成器T也存在着另一些优势. 首先，可以最大化后门图像生成器T在分类器f上的表现. 其次，由于本文将T建模为条件生成函数，这样使得触发模式能够根据输入图像的不同而产生变化，后门图像变得难以检测，从而进一步增强后门攻击的隐蔽性. 最后，整个触发模式的生成与选择也更加自动化，从而提升后门攻击的效率.

算法 1. 基于感知相似性的多目标优化隐蔽图像后门攻击方法.

输入：数据集 S，阶段 1 迭代次数 k，阶段 2 迭代次数 m，分类器f的学习率 γ_f，后门图像生成器 T 的学习率 γ_T，批尺寸大小 b；

输出：中毒分类器模型的参数ω^*，后门图像生成器模型的参数σ^*.

① 初始化 ω 和 σ；

/*阶段 1 训练*/

② $i\leftarrow 0$; /* i 表示当前迭代次数*/

③ repeat

④ 从 S 中提取样本 (x, y)；

⑤ $\sigma_{i+1}\leftarrow\sigma_i-\gamma_T\nabla_{\sigma_i}[\beta L(f_{\omega_i}(T_{\sigma_i}(x)),\eta(y))+\delta L(x,T_{\sigma_i}(x))]$；

⑥ $\omega_{i+1}\leftarrow\omega_i-\gamma_f\nabla_{\omega_i}[\alpha L(f_{\omega_i}(x),y)+(1-\alpha)L(f_{\omega_i}(T_{\sigma_{i+1}}(x)),\eta(y))]$；

⑦ $i\leftarrow i+1$；

⑧ until $i=k$；

⑨ end repeat；

⑩ $\sigma^*\leftarrow\sigma_k$；

/*阶段 2 训练*/

⑪ $i\leftarrow 0$, $\omega_0^k\leftarrow\omega_k$；

⑫ repeat

⑬ 从 S 中提取样本 (x, y)；

⑭ $\omega_{i+1}^k\leftarrow\omega_i^k-\gamma_f\nabla_{\omega_i^k}[\alpha L(f_{\omega_i^k}(x),y)+(1-\alpha)L(f_{\omega_i^k}(T_{\sigma^*}(x)),\eta(y))]$；

⑮ $i\leftarrow i+1$；

⑯ until $i=m$；

⑰ end repeat；

⑱ $\omega^*\leftarrow\omega_m^k$.

在阶段 1，首先对模型参数 σ 和后门图像生成器模型参数 ω 进行初始化. 然后，按照循环次数 k，从训练数据集 S 中随机抽样一个样本 (x, y). 根据中毒分类器模型在有毒数据集上的损失函数$L(f_{\omega_i}(T_{\sigma_i}(x)),\eta(y))$、后门图像与对应原始图像的感知损失 $L(x, T_{\sigma_i}(x))$等计算出梯度，更新 σ 的值. 根据中毒分类器在干净数据集上的损失$L\left(f_{\omega_i}(x),y\right)$、中毒分类器在有毒数据集上的分类损失$L(f_{\omega_i}(T_{\sigma_{i+1}}(x)),\eta(y))$等计算

出梯度，更新 ω 的值. 重复执行上述步骤，直到完成 k 次循环或者达到收敛条件. 阶段 2，需要将阶段 1 中得到的最新 ω_k 赋值为 ω_0^k，用作本阶段的起点. 同样地，按照循环次数 m，从训练数据集 S 中随机抽样一个样本 (x, y). 根据中毒分类器在干净数据集上的损失 $L\left(f_{\omega_i^k}(x), y\right)$、中毒分类器在有毒数据集上的分类损失 $L\left(f_{\omega_i^k}(T_{\sigma'}(x)), \eta(y)\right)$ 等计算出梯度，更新 ω^k 的值. 重复执行上述步骤，直到完成 m 次循环或者达到收敛条件.

3 实验与分析

3.1 实验准备

3.1.1 数据集

本文选择 4 个广泛使用的数据集进行隐蔽后门攻击研究：MNIST[21]，CIFAR-10[22]，GTSRB[23] 和 CelebA[24]. 由于 CelebA 数据集中每张图像有 40 个属性，不适合用作多分类. 因此，我们采取 Salem 等人[25] 的建议，选择了最平衡的 3 个属性，分别是浓妆、微张嘴和微笑，然后将它们组合起来，形成 8 个类别.

3.1.2 分类模型

为了构建图像数据分类器，我们按照文献 [12] 的建议对 CIFAR-10 和 GTSRB 数据集采用 Pre-activation Resnet-18（简写为 PreActRes18）模型；对于 CelebA 数据集采用 ResNet18 模型；对于 MNIST 数据集采用 CNN 模型. 具体情况如表 1 所示.

Table 1 Datasets and Classifiers Used in Our Experiments

表 1 本文实验中使用的数据集以及分类器

数据集	主体	类别	输入大小	训练数量	分类器
MNIST	手写数字	10	28×28×1	60 000	CNN
CIFAR-10	一般对象	10	32×32×3	50 000	PreActRes18
GTSRB	交通标志	43	32×32×3	39 252	PreActRes18
CelebA	脸部属性	8	64×64×3	202 599	ResNet18

3.1.3 基 线

对于攻击实验，我们将本文提出的方法与 WaNet[12] 和 LIRA[13] 攻击方法进行比较，WaNet 采取向原始图像添加精心设计的扭曲场来生成后门图像，生成的后门图像相比于传统方法更加自然，攻击成功率也明显优于先前的攻击. LIRA 采取添加隐蔽噪声的方式进行攻击，更加不易于发现.

3.1.4 攻击设置

在 Pytorch 环境下采取与对照组 WaNet 和 LIRA 相同的攻击设置，使用随机梯度下降 (stochastic gradient descent, SGD) 训练网络. 首先使用交替更新算法训练图像分类器和后门图像生成器 60 个周期. 然后将学习率设置为 0.01，使用后门图像生成器对分类器进行微调，每 100 个训练周期后学习率变为 1/10，训练 500 个周期直至网络收敛. 对于其它超参数，我们在 4 个数据集上设置隐蔽系数为 0.05、感知相似性损失倍数为 10 来保持隐蔽性. 通常隐蔽系数越大、感知性相似性损失倍数越小，后门攻击越容易成功.

3.1.5 性能评价指标

本文采用有毒子集在中毒模型上的攻击成功率（attack success rate, ASR）来评估本文方法 PMOA 的攻击性能，使用干净子集在中毒模型上的干净准确率（clean accuracy, CA）来评估植入后门功能后对分类模型原本性能是否有影响，并将这 2 种评价指标与基线实验相比较.

为了验证 PMOA 方法的隐蔽性，本文使用了 3 种图像质量评价指标来评价后门图像的保真度，包括学习感知图像块相似度（learned perceptual image patch similarity, LPIPS）、峰值信噪比（peak signal-to-noise ratio, PSNR）和结构相似性（structural similarity, SSIM）. 同时为了评估 PMOA 方法在实际生产环境中的隐蔽性效果，参照 WaNet 进行了人工检查实验.

3.2 攻击性能评价

3.2.1 多对 1（all-to-one）攻击

表 2 中的数据展示了 all-to-one 攻击设置下 3 种后门攻击方法在 4 个不同数据集上的干净准确率和攻击成功率. 通过向分类模型的训练样本中加入生成的后门数据，本文的攻击可以成功地向模型中植入后门功能. PMOA 在 4 个数据集的攻击成功率均达到了 100%. 在干净图像子集的测试中，与 WaNet 和 LIRA 相比，干净准确率没有下降.

Table 2 The Results in all-to-one Backdoor Attacks

表 2 all-to-one 后门攻击结果

数据集	WaNet		LIRA		PMOA	
	干净准确率	攻击成功率	干净准确率	攻击成功率	干净准确率	攻击成功率
MNIST	0.99	0.99	0.99	1.00	0.99	1.00
CIFAR-10	0.94	0.99	0.94	1.00	0.94	1.00
GTSRB	0.98	0.98	0.98	1.00	0.98	1.00
CelebA	0.78	0.99	0.78	1.00	0.78	1.00

3.2.2 多对多（all-to-all）攻击

表 3 中的数据展示了 all-to-all 攻击设置下的实验结果. 由于 all-to-all 攻击设置将后门标签设置为对

Table 3 The Results in all-to-all Backdoor Attacks

表 3 all-to-all 后门攻击结果

数据集	WaNet		LIRA		PMOA	
	干净准确率	攻击成功率	干净准确率	攻击成功率	干净准确率	攻击成功率
MNIST	0.99	0.95	0.99	0.99	0.99	0.99
CIFAR-10	0.94	0.93	0.94	0.94	0.94	0.94
GTSRB	0.99	0.98	0.99	0.99	0.99	0.99
CelebA	0.78	0.78	0.77	0.77	0.78	0.78

真实标签的一种偏移, 因此 all-to-all 设置下的后门攻击相比于 all-to-one 要更加困难. 但是 PMOA 在具有更好隐蔽性的情况下, 干净准确率没有下降. 其中 GTSRB 数据集的攻击成功率为 0.99, 高于 WaNet 方法的 0.98; CelebA 数据集的干净准确率和攻击成功率分别为 0.78 和 0.78, 均高于 LIRA 方法的 0.77.

3.3 隐蔽效果

图 5 和图 6 展示了后门图像中添加的噪声大小, 从中可以明显看出, PMOA 相比 LIRA 无论是在 all-to-one 还是 all-to-all 攻击策略中生成的原始噪声都显著减少, 取噪声的 1/20 作为触发模式后生成的后门图像具有更好的隐蔽性.

表 4 和表 5 展示了后门图像在 3 种图像质量评价指标方面的对比结果.可以看出, 无论是哪种数据集、哪种攻击方法, PMOA 的 LPIPS 值都远低于 LIRA, 且十分趋近于 0, 说明在视觉上 PMOA 生成的后门图像与原始图像区别及其微小, 人眼几乎无法分辨. 而且 PMOA 的 PSNR 与 SSIM 值均高于 LIRA, 说明 PMOA 生成的图像具有更好的保真度. 这都进一步表明了 PMOA 方法生成的后门图像具有极高的隐蔽性.

3.4 人工检查实验效果

从 GTSRB 数据集中随机选取干净图像, 并生成对应的后门图像, 最终形成用于人工检测的 100 张测试图像. 选取 8 名人员参与判定, 以区分哪些图像是原始图像, 哪些是携带触发模式的后门图像. 人工检查结果如表 6 所示, 该结果表明 PMOA 对参与者产生了强大的干扰能力, 导致他们很难分辨原始图像与对应的后门图像, 以至于在识别原始图像时的准确度接近于随机猜测. 这进一步表明了 PMOA 方法

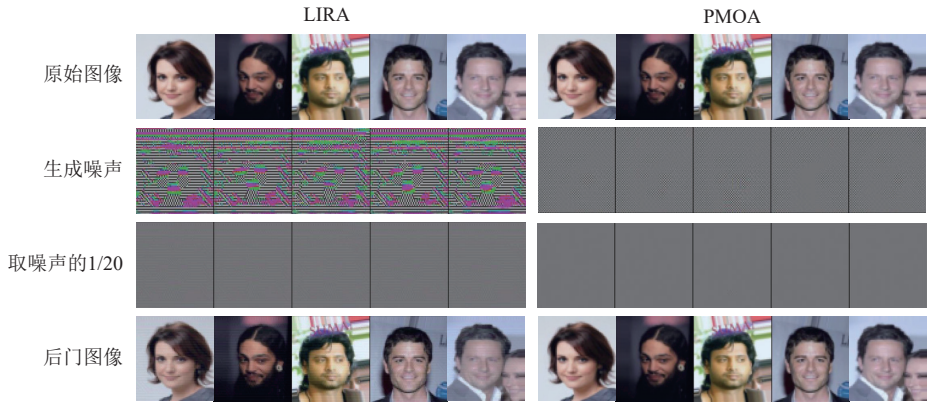

Fig. 5 Comparisons of backdoor images and noise under the all-to-one setting

图 5 all-to-one 设置下后门图像和噪声对比

Fig. 6 Comparison of backdoor images and noise under the all-to-all setting

图 6 all-to-all 设置下后门图像和噪声对比

Table 4 Comparison of Image Evaluation Metrics in all-to-one Backdoor Attacks

表 4 all-to-one 后门攻击中图像评价指标对比

数据集	LIRA			PMOA		
	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM
MNIST	0.000 039	51.632 35	0.975 38	0.000 026	53.269 35	0.986 17
CIFAR-10	0.000 160	52.114 51	0.998 46	0.000 003	62.720 48	0.999 83
GTSRB	0.000 123	45.787 00	0.988 01	0.000 079	45.922 23	0.996 94
CelebA	0.003 069	40.167 51	0.963 31	0.000 057	48.776 86	0.994 51

Table 5 Comparison of Image Evaluation Metrics in all-to-all Backdoor Attacks

表 5 all-to-all 后门攻击中图像评价指标对比

数据集	LIRA			PMOA		
	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM
MNIST	0.022 333	42.976 15	0.980 39	0.000 787	42.992 55	0.983 49
CIFAR-10	0.002 317	25.563 96	0.971 99	0.001 610	25.636 16	0.982 85
GTSRB	0.045 199	30.039 88	0.972 78	0.000 384	33.167 50	0.976 61
CelebA	0.000 737	46.096 07	0.996 90	0.000 007	59.674 44	0.999 73

Table 6 Manual Inspection Experiment for Accuracy

表 6 人工检查实验准确率

人员	准确率/%	人员	准确率/%
A	40	*E*	46
B	52	*F*	50
C	48	*G*	54
D	44	*H*	42

生成的后门图像具有很好的保真度，视觉上十分接近原始图像，突出了基于感知相似性后门攻击方法的隐蔽性.

4 结 论

本文主要围绕后门攻击图像隐蔽性问题，采用感知相似性与多目标优化方法，降低了后门图像与原始图像之间的差异，人眼难以分辨. 同时保持原始干净样本分类准确率不下降，最大化后门图像分类准确率.

为了实现这种方法，采取了一系列有效的措施. 首先，使用感知相似性损失函数来隐藏后门触发模式，降低了后门图像与对应原始图像在视觉上的差异. 其次，采用多目标优化方法解决了在被毒害模型上任务间冲突的问题，从而确保了模型性能不受影响. 最后，我们采取了两阶段的训练方法，提高触发模式的训练效率.

总之，本文提出的基于感知相似性的多目标优化隐蔽图像后门攻击方法具有很好的实用性和可行性，在提高深度学习模型的安全性和鲁棒性方向具有很好的参考价值. 未来，我们还需要进一步研究这种方法的适用范围和局限性，并不断探索新的解决方案，以应对不断出现的安全挑战.

作者贡献声明： 朱素霞对本文的构思、实验和论文撰写等提出了针对性的指导意见；王金印提出研究思路，设计实验，分析数据，并撰写论文；孙广路对设计思路、论文撰写等提供指导和审阅意见.

参 考 文 献

[1] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901

[2] Gu Tianyu, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain [J/OL]. (2019-05-11) [2023-06-19].https://arxiv.org/abs/1708.06733, 2019

[3] Gu Tianyu, Liu Kang, Dolan-Gavitt B, et al. BadNets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230−47244

[4] Yao Yuanshun, Li Huiying, Zheng Haitao, et al. Latent backdoor attacks on deep neural networks [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security (CCS'19). New York: ACM, 2019: 2041−2055

[5] Chen Dawei, Fu Anmin, Zhou Chunyi, et al. Federated learning backdoor attack scheme based on generative adversarial network[J]. Journal of Computer Research and Development, 2021, 58(11): 2364−2373(in Chinese)
(陈大卫，付安民，周纯毅，等. 基于生成式对抗网络的联邦学习后门攻击方案[J]. 计算机研究与发展，2021，58(11)：2364−2373)

[6] Zhang Quan, Ding Yifeng, Tian Yongqiang, et al. Advdoor: Adversarial backdoor attack of deep learning system [C] //Proc of the 30th ACM SIGSOFT Int Symp on Software Testing and Analysis. New York: ACM, 2021: 127−138

[7] Li Shaofeng, Xue Minhui, Zhao B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088−2105

[8] Quiring E, Rieck K. Backdooring and poisoning neural networks with image-scaling attacks [C] //Proc of 2020 IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2020: 41−47

[9] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks [C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11957−11965

[10] Wong E, Schmidt F, Kolter Z. Wasserstein adversarial examples via

projected sinkhorn iterations［J/OL］. (2019-02-21) ［2023-06-19］. https://arxiv.org/abs/1902.07906v2

［11］ Chen Xinyun, Liu Chang, Li Bo, et al. Targeted backdoor attacks on deep learning systems using data poisoning［J/OL］. (2017-12-15) ［2023-06-19］.https://arxiv.org/abs/ 1712.05526, 2017

［12］ Nguyen A, Tran A. WaNet−imperceptible warping-based backdoor attack［J/OL］. (2021-02-20) ［2023-06-19］.https://arxiv.org/abs/2102.10369, 2021

［13］ Doan K, Lao Yingjie, Zhao Weijie, et al. Lira: Learnable, imperceptible and robust backdoor attacks［C］//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 11966−11976

［14］ Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps［J/OL］. (2014-04-19) ［2023-06-19］.https://arxiv.org/abs/1312.6034, 2014

［15］ Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization［J/OL］. (2015-06-22) ［2023-06-19］.https://arxiv.org/abs/ 1506.06579, 2015

［16］ Justin J, Alexandre A, Li F. Perceptual losses for real-time style transfer and super-resolution［C］//Proc of the 14th European Conf on Computer Vision (ECCV2016). Berlin: Springer, 2016: 694-711

［17］ Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric［C］//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 586−595

［18］ Sener O, Koltun V. Multi-task learning as multi-objective optimization［C］//Proc of the 32nd Int Conf on Neural Information Processing Systems (NIPS'18). New York: ACM, 2018: 525−536

［19］ Désidéri J A. Multiple-gradient descent algorithm for multiobjective optimization［J］. Comptes Rendus Mathematique, 2012, 350(5/6): 313−318

［20］ Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks［J］. Communications of the ACM, 2020, 63(11): 139−144

［21］ Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition［J］. Proceedings of the IEEE, 1998, 86(11): 2278−2324

［22］ Krizhevsky A. Learning multiple layers of features from tiny images［J/OL］. (2019-04-08) ［2023-06-19］. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

［23］ Stallkamp J, Schlipsing M, Salmen J, et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition［J］. Neural Networks, 2012, 32: 323−332

［24］ Liu Ziwei, Luo Ping, Wang Xiaogang, et al. Deep learning face attributes in the wild［C］//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 3730−3738

［25］ Salem A, Wen R, Backes M, et al. Dynamic backdoor attacks against machine learning models［C］//Proc of the IEEE 7th European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2022: 703−718

Zhu Suxia, born in 1978. PhD, professor, PhD supervisor. Member of CCF. Her main research interests include privacy and security, IoT, and parallel computing.

朱素霞, 1978年生. 博士，教授，博士生导师.CCF会员. 主要研究方向为隐私与安全、物联网、并行计算.

Wang Jinyin, born in 1998. Master candidate. His main research interests include machine learning security and privacy preserving.

王金印, 1998年生. 硕士研究生. 主要研究方向为机器学习安全与隐私保护.

Sun Guanglu, born in 1979. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include artificial intelligence, network and information security, and intelligent information processing.

孙广路, 1979年生. 博士，教授，博士生导师.CCF杰出会员. 主要研究方向为人工智能、网络与信息安全、智能信息处理.