软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2400−2413 [doi: 10.13328/j.cnki.jos.006949]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于奇异值分解的隐式后门攻击方法*

吴尚锡, 尹雨阳, 宋思清, 陈观浩, 桑基韬, 于 剑

(北京交通大学 计算机与信息技术学院, 北京 100091)

通信作者: 桑基韬, E-mail: jtsang@bjtu.edu.cn

摘 要: 深度神经网络训练时可能会受到精心设计的后门攻击的影响. 后门攻击是一种通过在训练集中注入带有后门标志的数据, 从而实现在测试时控制模型输出的攻击方法. 被进攻的模型在干净的测试集上表现正常, 但在识别到后门标志后, 就会被误判为目标进攻类. 当下的后门攻击方式在视觉上的隐蔽性并不够强, 并且在进攻成功率上还有提升空间. 为了解决这些局限性, 提出基于奇异值分解的后门攻击方法. 所提方法有两种实现形式: 第 1 种方式是将图片的部分奇异值直接置零, 得到的图片有一定的压缩效果, 这可以作为有效的后门触发标志物. 第 2 种是把进攻目标类的奇异向量信息注入到图片的左右奇异向量中, 也能实现有效的后门进攻. 两种处理得到的后门的图片, 从视觉上来看和原图基本保持一致. 实验表明, 所提方法证明奇异值分解可以有效地利用在后门攻击算法中, 并且能在多个数据集上以非常高的成功率进攻神经网络.

关键词: 后门攻击; 奇异值分解; 进攻成功率; 隐蔽性

中图法分类号: TP18

中文引用格式: 吴尚锡, 尹雨阳, 宋思清, 陈观浩, 桑基韬, 于剑. 基于奇异值分解的隐式后门攻击方法. 软件学报, 2024, 35(5): 2400–2413. http://www.jos.org.cn/1000-9825/6949.htm

英文引用格式: Wu SX, Yin YY, Song SQ, Chen GH, Sang JT, Yu J. Stealthy Backdoor Attack Based on Singular Value Decomposition. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2400–2413 (in Chinese). http://www.jos.org.cn/1000-9825/6949.htm

Stealthy Backdoor Attack Based on Singular Value Decomposition

WU Shang-Xi, YIN Yu-Yang, SONG Si-Qing, CHEN Guan-Hao, SANG Ji-Tao, YU Jian

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100091, China)

Abstract: Deep neural networks can be affected by well-designed backdoor attacks during training. Such attacks are an attack method that controls the model output during tests by injecting data with backdoor labels into the training set. The attacked model performs normally on a clean test set but will be misclassified as the attack target class when the backdoor labels are recognized. The currently available backdoor attack methods have poor invisibility and are still expected to achieve a higher attack success rate. A backdoor attack method based on singular value decomposition is proposed to address the above limitations. The method proposed can be implemented in two ways: One is to directly set some singular values of the picture to zero, and the obtained picture is compressed to a certain extent and can be used as an effective backdoor triggering label. The other is to inject the singular vector information of the attack target class into the left and right singular vectors of the picture, which can also achieve an effective backdoor attack. The backdoor pictures obtained in the two kinds of processing ways are basically the same as the original picture from a visual point of view. According to the experiments, the proposed method proves that singular value decomposition can be effectively leveraged in backdoor attack algorithms to attack neural networks with considerably high success rates on multiple datasets.

Key words: backdoor attack; singular value decomposition; attack success rate; stealthy

---

* 基金项目: 国家自然科学基金 (61832002, 62172094); 北京市杰出青年基金 (JQ20023)
收稿时间: 2022-10-11; 修改时间: 2023-03-21; 采用时间: 2023-03-31; jos 在线出版时间: 2023-09-13
CNKI 网络首发时间: 2023-09-14

1 介 绍

随着深度学习方法逐渐成为机器学习的新一研究重点, 图像识别、自然语言处理、语音识别领域下的相关模型与深度学习技术结合后取得了惊人的进展[1−3], 许多任务在准确率、迁移性等评价指标上均获得了令人满意的结果, 并在实际应用场景中出色地表现了模型自身的实力. 但是深度学习模型的鲁棒性和可解释性仍是困扰学界的主要问题, 从对抗样本问题、数据投毒问题[4−6]到后门攻击问题, 无一不阻碍深度学习模型的部署和发展[7]. 其中后门攻击是一种新型的攻击技术, 在众多针对神经网络的攻击方法中, 后门攻击对神经网络的破坏力极大[8−13]. 后门攻击是指攻击者通过篡改训练数据, 让模型训练过程受到污染, 从而让攻击者可以通过可控的方式让污染的模型输出错误的结果的攻击方法. 目前学界对后门攻击的研究主要集中在以下两点: 首先, 在准确性方面, 学界看重模型植入后门之后接受特定输入数据后成功进攻的比率; 其次, 在隐蔽性方面, 学界希望用户难以发觉样本上携带的特殊标记[14,15].

当下的后门攻击的模式, 或是通过增加特殊像素点[4], 或通过增加条纹信息[16], 或是利用物反射这类物理方式加入后门信息[5]. 但这些模式的最大弊病在于他们将后门的信息都注入在了像素实空间上, 在视觉表现上具有比较明显的异样特征, 如果人类来辨别的话, 还是能够非常轻松地识别出后门的存在.

为了解决上述问题, 我们提出基于奇异值分解的后门方法 (SVD backdoor attack), 试图不在像素的实空间上进行后门的植入, 而是把图像进行奇异值分解, 在奇异向量或奇异值中注入带有目标进攻类的后门进攻信息, 从而避免视觉上的巨大差异. 我们将其细化成了两种具体可操作的方法. 第 1 种方法是在奇异值上设置后门标志, 称之为奇异值后门进攻 (singular-value backdoor attack). 由于每组奇异值和奇异向量都包含着图像的不同信息, 我们将图像的较靠后的奇异值全部置 0, 得到的图像有一定的压缩效果, 但是这样缺省奇异值的信息也能作为后门触发的标志. 第 2 种方法是在左右奇异向量上设置后门标志, 我们称之为奇异向量后门进攻 (singular-vector backdoor attack). 具体思路是将目标进攻类的奇异向量信息注入到干净的图片上, 得到后门进攻训练数据集, 并在测试集上的图片以注入同样的奇异向量来作为后门触发的标志. 基于奇异值分解的后门进攻方法存在着几点优势. 首先由于后门触发标志被注入在了部分奇异向量或者奇异值中, 再进行奇异值合成之后得到的新图像, 新图像有一定的压缩或者轻微扰动. 但在人眼上, 后门进攻后的图片和原图的差异几乎不存在. 其次是奇异值信息能够被神经网络捕捉到, 因此在进攻成功率上也有非常不错的表现.

本文的主要贡献如下.

(1) 我们提出了基于奇异值分解的后门攻击方法, 将后门标志信息注入到图片的奇异向量或者奇异值中, 并且证明了该后门进攻方法具有较强的隐蔽性.

(2) 在投毒比例相同、网络结构相同的情况下, 对比主流的后门进攻方法, 我们的方法在进攻成功率、进攻隐蔽性上表现都更好.

2 相关工作

当前, 后门攻击领域的研究可分为探索性后门攻击和推理式后门攻击. 前者通过输入被攻击方精心设计的数据, 影响深度学习模型的输出结果, 达到攻击目的; 后者则通过改变神经网络的训练过程, 降低神经网络输出结果的准确率, 达到攻击目的.

Gu 等人提出的 BadNet 首次系统地研究了深度学习领域的后门攻击问题[4]. 并通过修改标签的方式对输入数据进行调整, 在 MNIST 数据集上实现了后门攻击. 然而, 这种方法虽然能够极大破坏模型的准确性, 却需要针对每批输入的数据进行投毒, 且必须借助模型重新训练环节确保后门已被植入. 为克服这一局限性, Tang 等人[17]设计了一种无需训练的后门攻击方法. 他们提出的方法将一个小型木马模块插入到模型中, 让木马模块降低模型图像处理的效果. 这一方法不仅可适用于不同的深度学习系统, 而且不会降低良性样本预测时的准确率. 然而, 这一方法也有自身的局限: 附加于原模型的额外结构容易被检测出来, 隐蔽性较差. 为提高植入模型的后门的隐蔽性, Liu 等人设计了基于物理反射的后门攻击方案[5]. 在这一方案中, 投毒的数据由正常图像与反射影像结合生成, 后

门则被设计为物体的反射成像特征. 这一方法由于避开了对训练数据及其标签的可以修改并躲避了基于 input-filtering 思想设计的防御策略, 兼具攻击性和一定程度的隐蔽性, 后门攻击效果显著. 为进一步提升后门攻击方法的隐蔽性, Li 等人设计了 PASS 等评价指标, 并通过正则化方式使后门攻击不易被人类察觉[18]. Turner 等人通过干扰图像像素进而植入后门, 而非对图像进行替换[19]. Nyuyen 等人设计了 WaNet, 其功能是通过几何变换实现了原图像的翘曲, 进而完成后门进攻[20]. 图像缩放也逐渐成为另一植入模型的攻击方式[21,22]. Wenger 等人综合了物理变换, 并据此设计了相应的后门攻击方案[23]. Bagdasaryan 等人将后门攻击视为一种特殊的多任务学习, 并设计了针对损失函数计算进行攻击的后门攻击方案[24]. Shumailov 等人通过更改训练顺序设计了后门攻击方案[25]. Kurita 等人则设计了根据调整不同的投毒权重设计了后门进攻方案[26].

伴随着后门攻击领域的研究逐渐深入, 针对后门攻击的防御策略也逐渐受到人们关注[27−34], 这使得对于神经网络的攻击难度逐渐上升. 传统的基于标签翻转修改、图像几何变化的攻击手段由于模型防御技术发展正逐渐失效. 为进一步提升攻击的成功率, Chen 等人在极少部分的输入数据中混入了噪声图像, 譬如在原图中引入 Hello Kitty 的图案、为图中人物佩戴由攻击方设计的眼镜, 并通过调节混合程度进行后门攻击[35], 如图 1. 这一方法可在对数据进行少量修改的前提下实现较好的后门攻击. Bagdasaryan 等人利用联邦学习的聚合机制, 在局部模型最终聚合的过程中将中毒的数据提交给聚合器, 以此向原模型植入后门[36]; Chen 等人在此基础上设计了对联邦元学习的后门攻击方法[37]. Zhu 等人针对模型迁移性设计了后门攻击模型, 他们将样本发布在网上, 假借权威机构或者研究员下载后将对这些事实上中毒的样本进行标记完成后门的植入[38]. 受此方法启发, Souri 等人[39]和 Liu 等人[40]将后门攻击转化为双层优化问题, 并分别在此基础上提出了新的后门攻击方法. Garg 等人则通过更改决策边界的方式植入后门, 据此设计后门进攻方案[41].

(a) Airplane (b) Speed limit (c) Cat (d) Woman

图 1 常见后门攻击标志物样式

在当前后门攻击领域中, 人们不仅关注后门进攻成功率, 也正在探索具有更强隐蔽性的后门进攻方法. 继 Gu 等人[4]和 Chen 等人[35]分别测试比较后门进攻的隐蔽性后, Liu 等人开始观察不同后门进攻成功率下后门攻击方法隐蔽性[5]. Zhang 等人通过提取图像边缘信息作为后门的标志, 采用注射网络生成了受到投毒的图像数据[42]. Li 等人则通过攻击目前防御模型体系中对后门攻击的假设前提设计出了具有样本特定触发器的后门攻击方法[43]. Yao 等人打破了传统后门攻击的模式, 让进攻方按照自身需求对预训练模型进行训练, 然后将预训练模型发布在网上, 这些后门在用户下载该预训练模型的过程中植入了用户的模型[44]. 这一模型不仅避开了传统防御模型对于相关标签的防御, 而且扩展性较强. Tan 等人为了使植入的后门不被防御模型中的传统检测算法识别, 通过设计自适应的训练算法来优化模型原始的损失函数, 进而建立了具有旁路性质的后门[45]. Zhong 等人[46]借鉴了全局对抗扰动的思路[47], 基于 L2 范数最小化对扰动进攻进行优化, 提升了后门进攻的隐蔽性. 此后, 基于 Lp 范数优化进攻结果的方法陆续被提出, 这些成果启发人们在后门进攻隐蔽性进一步地研究[48,49].

3 方 法

3.1 奇异值分解

奇异值分解广泛应用于深度学习领域, 它不仅可以用于图像处理中的图像降噪、数据压缩, 还可以用于推荐系统、自然语言等多个领域[50]. 奇异值分解类似于矩阵的特征分解, 但并不要求分解的矩阵为方阵. 假设矩阵 A 为 $m\times n$ 的矩阵, 那么奇异值分解可定义为:

$$A_{m\times n}=U_{m\times m}\Sigma_{m\times n}V^{\mathrm{T}}_{n\times n} \tag{1}$$

其中, U, V^{T} 均为可逆矩阵, 分别称为矩阵 A 的左奇异向量和右奇异向量. Σ 除了对角线外的元素都是 0, 对角线上的元素称为奇异值, 奇异值按照大小排列且快速衰减, 前 10% 甚至 1% 的奇异值的和就占全部的奇异值之和的 99% 以上, 所以我们可以用前 k (k 远小于 m 和 n) 个奇异值来近似描述矩阵. 奇异值分解往往能解析矩阵中隐含的重要信息, 且重要性和奇异值大小有一点相关性.

3.2 后门攻击问题定义

给定 P 类图像数据集 $D=\{(x_i,y_i)\}_{i=0}^{n}$, 其中 $x\in X\subset R^d$ 表示 d 维输入空间中的样本, $y\in Y=\{1,2,\ldots,P\}$ 表示图像的真实标签. 我们假设受害者将黑盒模型的训练外包给投毒者, 因此投毒者可以访问目标模型的训练数据集 D_{train}. 后门攻击的过程可以表示为对给定的图像数据 x_i, 进行了 $x_i^b=x_i+v$ 操作, 其中 v 代表后门的触发标志, 并将原标签 y_i 修改为目标进攻标签 y_{target}, 将这些进行加入过后门的标志物的数据集定义为 D_{backdoor}. 投毒者修改一部分的训练集 D_{backdoor} 以达到后门攻击的目的, D_{backdoor} 与 D_{train} 的比例即为后门进攻比例. 用包含 D_{backdoor} 的训练集训练模型, 在投毒者把训练好的模型提交给受害者时, 如果训练好的被投毒模型与非投毒模型在测试集 D_{test} 上的准确率相接近, 受害者就会接受该模型.

3.3 基于奇异值分解的后门攻击方法

针对投毒有效性以及隐蔽性的综合考虑, 我们提出基于奇异值分解的后门进攻 (SVD backdoor attack). 奇异值分解提供了一种分析图片矩阵特征信息的手段, 我们针对奇异值分解得到的奇异矩阵作出攻击, 嵌入基于奇异值的后门. 嵌入后门后的中毒图片与原始图片差别较小, 且只需要少量的投毒比例就可以训练出一个强大的后门触发器, 该后门触发器在遇到包含特定奇异向量的图片时, 会将其识别为特定的类别, 并且在原始分类任务上表现良好.

SVD backdoor attack 的框架如图 2 演示, 其中细分成了两种可实行的办法.

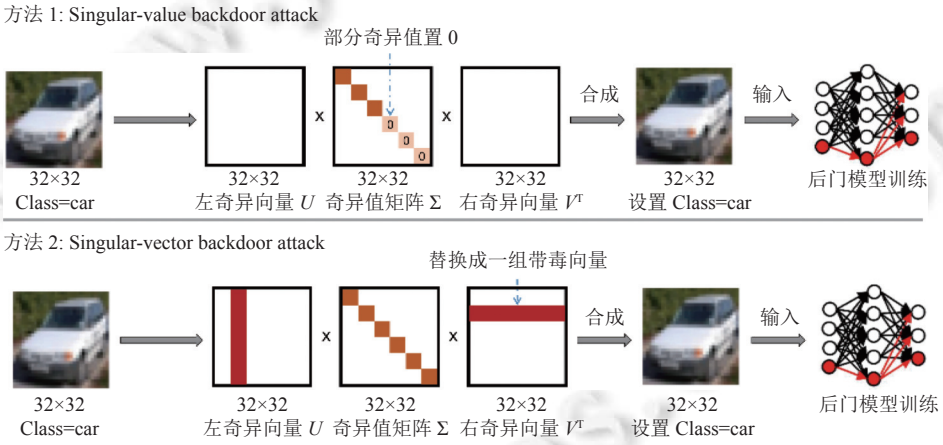


图 2 SVD backdoor attack 框架图

3.3.1 方法 1: 奇异值后门进攻 (singular-value backdoor attack)

方法 1 的思路是将图片靠后的一部分奇异值直接置 0, 因此来作为后门攻击的触发标志. 这部分丢弃的奇异值信息会作为后门的特征来作为目标攻击类的特征信息.

具体的操作如图 3 所示, 以 CIFAR10[18]为例子, 先对输入的干净图像 x_i 进行奇异值分解, 将会得到左右奇异值向量 U, V, 它们分别是 32×32 的酉矩阵和一个 32×32 半正定对角矩阵 Σ, 且其中只有对角线中存在非零元素. 然后将分解得到的 Σ 后 k 个值之后的奇异值置 0, 可以表示为:

$$x_i=U_i\Sigma_iV_i^{\mathrm{T}} \tag{2}$$

$$\Sigma_i[k:]:=0 \tag{3}$$

其中, :=代表将右边的值赋给左边, $\Sigma_i[k{:}]$ 代表第 k 个奇异值到最后一个奇异值, 将他们全部置为 0. 并且需要将原标签 y_i 修改为目标进攻标签 y_t . 按照上述的做法, 我们做出了奇异值进攻处理之后的图片效果如图 4 所示.

图 3　后门攻击示意图

(a) 未经处理的原图　　(b) Singular-value backdoor attack 处理后

图 4　Singular-value backdoor attack 进攻后的图片和原图对比

由于奇异值保存了图像的大量信息, 丢弃部分奇异值实质上是对图像进行了一个压缩的操作. 在我们的奇异值后门进攻中将 k 值设置的较为靠后的话, 则可以实现压缩的幅度比较小, 在人眼看起来几乎和原图一致, 但是同时又能起到后门的触发标志的作用.

3.3.2　方法 2: 奇异向量后门进攻 (singular-vector backdoor attack)

方法 2 的思路是将后门标志信息注入在了奇异向量中, 其中标志信息采用了目标进攻类的一组奇异向量. 神经网络会把标志信息作为分类输入图片为目标类的重要判断依据.

具体的操作是选取一个目标进攻类图片 x_t , 将其奇异值分解得到的奇异值向量作为后门的触发标志, 并将其注入到图像 x_i 中. 用数学表达式可以写成如下的形式:

$$x_i = U_i \Sigma_i V_i^{\mathrm{T}} \tag{4}$$

$$U_i'[:,k] := U_t[:,k] \tag{5}$$

$$V_i'[k,:] := V_t^{\mathrm{T}}[k,:] \tag{6}$$

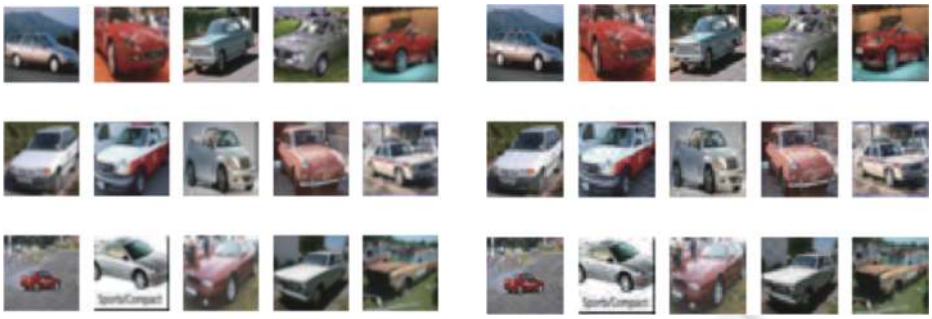
$$x_i' = U_i' \Sigma_i V_i'^{\mathrm{T}} \tag{7}$$

其中, $U_t[:,k]$ 代表目标进攻类 x_t 的 U 向量中第 k 列向量, $V_t^{\mathrm{T}}[k,:]$ 代表目标进攻类 x_t 的 V^{T} 向量中第 k 行向量, U' 代表注入后门之后的原图的 U 向量, x_i' 代表被进攻后的图像.

并且需要将原标签 y_i 修改为目标进攻标签 y_t . 按照上述的做法我们做出了 singular-vector backdoor attack 处理之后的图片效果如图 5 所示.

从图 5 中可以看出原图和进过奇异向量后门进攻之后的图片在肉眼上几乎看不出任何差距, 可以认为具有较好的隐蔽性.

相较于大多数后门攻击技术, 已有的技术往往是可见并且静态的. 即意味着对于不同的图片, 后门的标志分布的位置, 分布的样式是固定的, 是在相同的空间域触发后门标志. 而我们的操作可以认为是在奇异值和奇异向量领域上为图片增加后门, 因为奇异值和奇异向量涉及整张图片的信息, 所以加入的后门会在进过奇异值合成后会散布于整个图片中, 不会在某一区域集中体现, 这样从视觉上能够起到隐藏的效果.



(a) 未经处理的原图 (b) Singular-vector backdoor attack 处理后

图 5 Singular-vector backdoor attack 进攻后的图片和原图对比

4 实验分析

4.1 实验参数

(1) 数据集

在已有的后门攻击研究中, 通常使用的数据集是 MNIST、CIFAR10、CIFAR100 和 GTSRB[18]. 在下面的实验中我们主要以 CIFAR10 为例. CIFAR10 是一个十分类的数据集, 图像大小为 32×32, 总共有 6 万张图片, 其中训练集 5 万张, 测试集 1 万张.

(2) 网络结构

我们选取了几个比较主流的模型用于实验, 下面的实验中主要使用了 ResNet18[51]. 其中 epoch 设置为 12. 正常训练下 ResNet18 在 CIFAR10 的表现能够在测试集上达到 84% 左右的准确率.

(3) 评价指标

我们使用干净数据集准确性下降率 (clean accuracy drop, CAD) 来评估后门模型对干净测试数据集的影响. CAD 值越接近于 0, 代表对未加后门的数据影响越小. 意味着用户通过第三方训练得到的模型, 在使用正常的图片进行分类任务的时候, 不会展现出和纯净的模型有任何的差别, 这样训练得到的后门模型也不容易被发现.

采用进攻成功率 (attack success rate, *ASR*) 来评估后门的有效性. 进攻成功率可以表示为加了后门标志之后, 被分类为目标进攻类别的图像的比率.

用来衡量后门进攻后图片的隐蔽性的指标主要有 3 个, 包括了峰值信噪比、结构相似性指数、学习感知图像块相似度.

峰值信噪比 (*PSNR*)[18], 是最普遍, 最广泛使用的评鉴画质的客观量测法, 它的基本原理是用均方差 (MSE) 来衡量新图像和旧图像的差异, *PSNR* 的值越大, 代表图像失真的越小. 一般的基准是 *PSNR* 大于 40 dB, 认为图像的品质是极好的, 30–40 dB 图像质量是好的, 20 dB 以下的图像不可以接受. 我们认为 *PSNR* 值越大认为和原图的差距越小, 这样的后门标志隐蔽性越强, 因此我们用 *PSNR* 来衡量后门的隐蔽性.

结构相似性指数 (structural similarity index measure, SSIM) 是用于量化两幅图像间的结构相似性的指标[52]. SSIM 从亮度、对比度以及结构量化图像的属性, 用均值估计亮度, 方差估计对比度, 协方差估计结构相似程度. SSIM 值的范围为 0 至 1, 越大代表图像越相似. 如果两张图片完全一样时, SSIM 值为 1.

学习感知图像块相似度 (learned perceptual image patch similarity, LPIPS) 也称为“感知损失”, 是用于度量两张图像之间的差别[53]. LPIPS 更符合人类的感知情况, 其值越低表示两张图像越相似, 反之, 则差异越大.

4.2 奇异值保留个数和性能的关系

在奇异值后门进攻中, 主要的操作是把奇异值矩阵 Σ 的第 k 个之后的奇异值全部丢弃, 也就是说只保留前 k 个特征值信息, 以此来作为后门的标志.

图 6 作出了保留不同奇异值个数的图像效果, 比如左上角第 1 张就代表了只保留了一个奇异值的效果, 发现图片被极大地压缩, 变成了肉眼不可判断的模糊图片. 而随着保留的奇异值个数越来越多, 图片的清晰度越高, 越来越逼近真实图像. 在保留了约前 7 个特征值信息之后, 得到的图片基本和原图不再有肉眼可视上的差别.

图 6 保留不同奇异值个数对图像效果的影响

我们可以猜测, 如果左上角, 即只保留了 1 个奇异值的图片, 作为后门的标志, 那效果肯定是极佳的, 因为神经网络在学习的过程中凡是识别到极大模糊的图片都是被分为目标进攻类, 那么在测试集上的泛化性肯定也是非常好的. 但是这样的后门标志过于明显, 所以我们希望得到的结果是后门进攻成功率 (*ASR*) 和后门隐蔽性 (和原图的差异) 之间的一种权衡效果. 因此我们对保留不同奇异值数量的后门模型进行训练, 其中采用的模型是 ResNet18, 数据集是 CIFAR10, 设置的后门比例为 3%. 得到了如图 7 的结果.

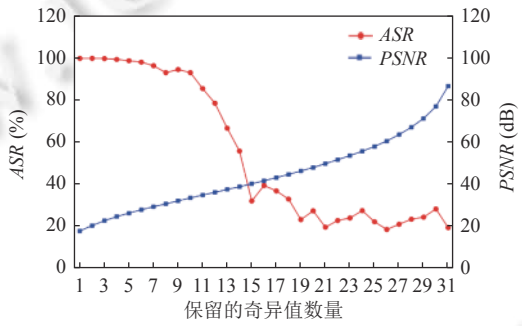


图 7 保留奇异值数量和 *ASR*、*PSNR* 的关系

从图 7 中可以看到随着保留奇异值数量的越来越多, 得到后门进攻成功率 *ASR* 也是越来越差. 产生这样的情况原因是随着图像越来越逼近原图, 经过后门处理后的图像和干净的图像很难再做出一个区分, 因此后门的效果会越来越差.

此外测量了模型的 CAD, 标准 ResNet18 在 CIFAR10 上的准确率在 85% 左右, 而我们在注入后门之后的模型, 在干净数据集上的准确性与标准模型的误差在±1% 左右, 基本上可以认为后门的注入对无后门的测试用例不产生任何影响.

由于后门进攻成功率 (*ASR*) 和后门的隐蔽性 (*PSNR*) 存在负相关的现象, 不存在成功率和隐蔽性都是最好的方法, 需要引入一个新的指标来刻画隐蔽性加上成功率, 因此我们定义了后门综合能力 (backdoor comprehensive ability, *BCA*) 来表示两者的关系, 数学表示为:

$$BCA = \alpha \cdot ASR + (1-\alpha) \cdot PSNR \tag{8}$$

其中, α 为比例常数, 用来代表对不同要素的重视程度, 如果在有些场景中后门进攻成功率相比隐蔽性更重要, 则增大 α 的取值. *ASR* 和 *PSNR* 的值需要进行归一化, 才可以进行叠加计算. *BCA* 的值越大代表了后门的能力越强.

根据上述的衡量方法, 我们做出了保留不同奇异值数量和后门综合能力的关系, 结果如图 8.

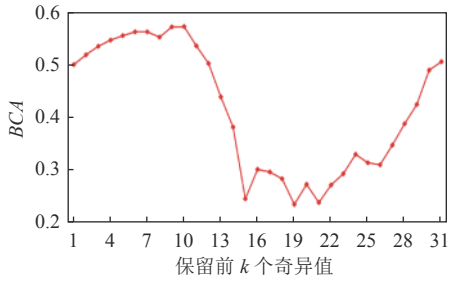


图 8　保留奇异值数量和 *BCA* 的关系

我们选取 *BCA* 值最大的作为最优的超参数设定值, 在 CIFAR10 数据集的例子中, 我们的给出的建议是保留前 10 个奇异值信息, 也就是说将 $\Sigma[10:32]$ 置为 0.

4.3　奇异向量注入位置和性能的关系

在方法 2 中, 我们替换第 k 个奇异值对应的奇异向量信息, 以此来作为后门的标志. 得到的图像效果如图 9 所示.

图 9　奇异向量注入位置对图像效果的影响

图 9 中 20 张子图片分别代表了替换第 k 个奇异值对应的奇异向量的效果. 在替换第 1 个奇异向量后, 由于奇异值权重很大, 产生对抗扰动会非常明显. 在替换第 5 个及以后的特征向量时, 得到的图片基本和原图不再有肉眼可视上的差别.

同样是量化了后门进攻成功率 (*ASR*) 和后门隐蔽性 (*PSNR*) 与参数的关系. 我们对数据集进行替换不同位置奇异向量的后门进攻, 参数设置同上, 得到了如图 10 的结果.

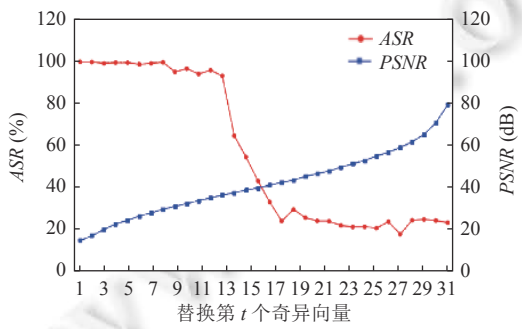


图 10　替换不同位置的特征向量与 *ASR*、*PSNR* 的关系

可以发现, 在替换 12 奇异向量之前的结果得到的进攻成功率基本都维持在了 95% 以上, 再往后的性能会有一个快速下降的情况. *PSNR* 的情况和方法 1 基本一致, 是随着注入奇异向量位置的靠后对图像影响逐步减小带来的 *PSNR* 的上升.

并且模型的 CAD 同样表现优秀, 基本上可以认为后门的注入对无后门的测试用例不产生任何影响. 为了权衡 *ASR* 与 *PSNR*, 找到最合适的进攻位置, 我们利用 *BCA* 分析, 得到了如图 11 的效果.

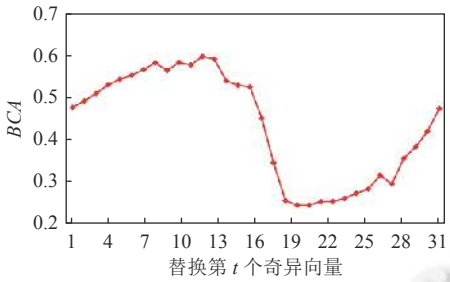


图 11 替换不同位置的特征向量与 *BCA* 的关系

我们选取 *BCA* 值最大的作为最优的超参数设定值, 在 CIFAR10 数据集的例子中, 我们的给出的建议是替换第 12 个奇异向量.

4.4 后门进攻比例和进攻成功率的关系

后门的比例对于后门攻击模型的影响重大. 因为如果后门设置的比例比较小时, 意味着不容易被发现, 但进攻的有效性可能会受到影响, 因此我们的实验量化了进攻成功率和投毒比例之间的关系, 其中参数选择为上述所描述的最优参数, 做出了曲线图.

由图 12 的显示结果可以知道, 后门进攻的成功率和后门设置的比例基本是成正相关的, 也就是说后门的比例越大, 后门的成功率越好. 因为更多的后门案例, 能够在分类器中学习到更多的关联特征, 因此在泛化时有着更加优秀的表现.

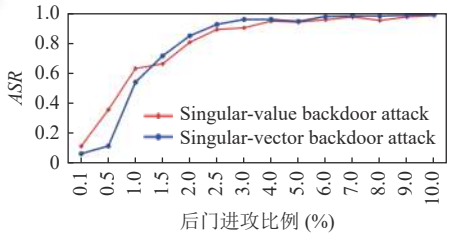


图 12 后门进攻比例和 *ASR* 的关系

从我们的结果发现, 当后门进攻比例大于 3% 之后, *ASR* 的值也超过了 90%, 这是一个较为满意的结果, 表明在小数据的后门攻击, 效果也能比较显著. 在进攻比例大于 6% 之后, 后门进攻成功率也基本为 98% 以上, 在接近 10% 左右的进攻比例, 则基本实现了 100% 的进攻成功率.

4.5 方法在不同数据集上的表现

为了验证方法的有效性, 我们选取了常用的数据集进行了进攻方法的测试, 主要选取了 MNIST、CIFAR10、CIFAR100、GTSRB. 采用的模型是 ResNet18, 通过对不同后门比例的设置, 测试了几个关键指标上的性能, 在表中相同配置下最优性能指标被加粗.

从表 1 和表 2 中可以看出, 奇异值后门进攻 (singular-value backdoor attack) 和奇异向量后门进攻 (singular-vector backdoor attack) 在 4 个数据集上均有较高的进攻成功率 (*ASR*), 在后门进攻比例超过 3% 时, 除了 CIFAR10 以外, 各数据集都实现了超过 99% 的后门进攻成功率. 在干净数据集准确性下降率 (CAD) 上均在 3% 以内, 表示在多数据集上我们的方法对干净数据集准确性影响较小. 在 MNIST、GTSRB 上的 CAD 接近于 0, 可以认为是对准确率毫无影响. MNIST 由于像素块较少, 细微的变动都会对 *PSNR* 造成巨大的影响, 但在另外 3 个数据集上,

PSNR 都超过了 30 dB, 可以认为图像失真程度低, 和原图差异不大. 学习感知图像块相似度 (LPIPS) 和结构相似性指数 (SSIM) 两个指标也同样用来刻画隐蔽性效果, LPIPS 基本接近 0, SSIM 接近于 1, 都代表了图片和原图差异细微.

表 1 奇异值后门进攻在不同数据集上的表现

数据集	后门比例	CAD (%)	*ASR* (%) ↑	*PSNR* (dB) ↑	LPIPS ↓	SSIM ↑
MNIST	0.01	**0.00**	92.60	26.62	0.0862	0.8502
	0.03	0.03	95.59			
	0.05	−0.04	**96.83**			
CIFAR10	0.01	**1.22**	95.00	30.36	0.0450	0.9798
	0.03	1.34	97.10			
	0.05	2.15	**98.60**			
CIFAR100	0.01	−0.35	76.00	34.96	0.0390	0.9799
	0.03	0.66	94.20			
	0.05	**−0.31**	**96.20**			
GTSRB	0.003	0.11	66.72	41.70	0.0127	0.9896
	0.004	0.07	94.32			
	0.01	**0.04**	**100.00**			
	0.03	0.05	**100.00**			

表 2 奇异向量后门进攻在不同数据集上的表现

数据集	后门比例	CAD (%)	*ASR* (%) ↑	*PSNR* (dB) ↑	LPIPS ↓	SSIM ↑
MNIST	0.01	**0.00**	95.68	25.54	0.0142	0.9931
	0.03	0.02	99.82			
	0.05	−0.01	**99.91**			
CIFAR10	0.01	0.43	85.40	35.06	0.0208	0.9916
	0.03	0.52	97.10			
	0.05	**0.32**	**98.10**			
CIFAR100	0.01	0.59	92.00	38.98	0.0132	0.9906
	0.03	**−0.17**	**99.67**			
	0.05	1.13	99.60			
GTSRB	0.003	0.10	44.56	42.10	0.0046	0.9961
	0.004	0.08	92.31			
	0.01	**0.03**	**100.00**			
	0.03	0.06	**100.00**			

4.6 现有后门攻击方法对比

我们用 SVD backdoor attack 方法和已有的常用后门攻击模型进行对比, 主要比较隐蔽性、进攻成功率、后门比例、CAD 等几个主要参数.

图 13 所示演示了几种主要的后门攻击方法, BadNet[4]是后门攻击的开山之作, 主要在右下角加了小方块来作为后门信息; SIG[16]方法主要在图片上叠加了条纹信息; Blend[35]的方法是把图片和其他的图片进行混合, 以混合的图片信息作为后门的标志.

(a) 原图 (b) BadNet (c) SIG (d) Blend (e) Singular-vector (f) Singular-value

图 13 不同后门攻击方法的图片效果

首先从视觉上来看, 目前主流的几种方法都有较为明显的可分辨性. 对于人类来说, 能够轻而易举地识别出异样信息. 这样的情况对于后门攻击模型来说是不可以接受的, 会极大破坏了隐蔽性的效果, 虽然已有的研究都试图降低视觉上的差异, 但始终存在在视觉域上的可见性.

而我们的 SVD backdoor attack 方法则是以奇异值信息来作为后门攻击的标志, 奇异值的舍弃对图像产生了一部分压缩的影响, 但这样的压缩在人眼的视觉上是很难分辨的.

从表 3 中数据可以分析得到, 当下主流的后门进攻模型的 CAD 差异都不大, 和原始的准确性差异较小, 都属于可以接受的范围.

表 3 不同后门攻击模型的性能对比

指标	比例	BadNet	Blend	SIG	Refool	SPM	Poison ink	Singular-value	Singular-vector
CDA (%)	3	1.62	0.11	0.26	0.80	**0.11**	0.35	0.43	1.22
	5	1.87	0.40	0.36	0.84	**0.10**	0.31	0.52	1.34
	10	0.50	**0.23**	0.55	0.20	0.93	0.53	**0.32**	2.15
ASR (%)	3	66.55	89.39	87.16	87.16	58.53	94.22	97.10	**97.12**
	5	65.36	90.99	89.79	89.79	57.69	93.58	98.11	**98.60**
	10	79.30	93.11	92.80	92.80	57.33	93.67	**98.40**	98.31
PSNR (dB) ↑	—	26.27	19.87	25.12	19.38	35.94	**42.95**	35.06	30.38

PSNR 上的数据可以衡量隐蔽性, 我们的两种方法 singular-value backdoor attack 和 singular-vector backdoor attack 方法的 *PSNR* 值能够达到 30 dB 以上, 按照标准来说属于和原图差异较小, 而已有的几种典型方法中 *PSNR* 值均在 30 dB 以下. Refool 和 Poison ink 是试图提升隐蔽性的两种进攻方法, 实验数据取自原论文, 从数值上可以看出 Refool 的隐蔽性和我们的相差较大. Poison ink 虽然在 *PSNR* 上超过了 40 dB, 但在 *ASR* 上和我们的方法差距过大. 如果我们对方法进行调参, 在同样的 *ASR* 情况下, 我们方法的 *PSNR* 值会更高. 我们认为在 *PSNR* 大于 30 dB 时, 隐蔽性已经很强, 此时增大 *ASR* 更为重要.

从进攻成功率上来衡量, 我们的两种方法在同比例下 *ASR* 上数值大幅超过了之前的模型, 可见我们的方法不仅在隐蔽性上有突破, 并且在成功率上也得到了显著提升.

4.7 消融实验

先前的实验结果都是基于 ResNet 结构的 backbone 进行的, 为了证明我们的方法对不同的 backbone 都有很强的进攻效果. 我们设定投毒率为 0.05, 在数据集 MNIST, CIFAR10, CIFAR100 和 GTSRB 数据集上进行实验, 并且分别在 VGG16[54], InceptionNet v4[55], ViT[56]结构的 backbone 上进行了我们设计的后门攻击, 效果如表 4 所示.

表 4 不同 backbone 下后门攻击的性能对比

backbone	注入率	*ASR* (%)			
		MNIST	CIFAR10	CIFAR100 ↑	GTSRB ↑
ResNet18	0.01	95.68	85.40	92.00	**100.00**
	0.03	99.82	97.10	**99.67**	**100.00**
	0.05	**99.91**	**98.10**	99.60	**100.00**
VGG16	0.01	95.74	85.05	92.11	**100.00**
	0.03	99.90	97.73	98.80	**100.00**
	0.05	**99.95**	**98.63**	**98.91**	**100.00**
InceptionNet v4	0.01	95.57	85.05	91.21	**100.00**
	0.03	**100.0**	96.45	98.57	**100.00**
	0.05	**100.0**	**97.57**	**98.79**	**100.00**
ViT16	0.01	96.62	84.88	90.33	**100.00**
	0.03	99.53	96.40	99.05	**100.00**
	0.05	**99.84**	**98.08**	**99.21**	**100.00**

从表 4 中可以看出, 我们设计的后门攻击方法对于不同的 backbone 结构, 都能很好地进行攻击, 说明我们的方法对于不同 backbone 并不敏感. 尤其是对于 ViT 这种基于注意力的 backbone 也能产生较好的攻击效果.

5 总 结

在本文中, 我们通过对图片奇异值分解进行全面的分析, 创新地提出了一种基于奇异值的后门进攻方法. 首先, 我们展示了如何利用奇异值分解将不可见并且强大后门嵌入到整张图片中. 通过大量实验, 我们得到的峰值信噪比以及后门成功率等多个指标展示了该进攻方法在隐蔽性和攻击效率方面的能力. 我们希望本文提出的意见能够激发对后门攻击与奇异值领域之间关系的更深入的研究.

References:

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[2] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 2011, 12(12): 2493–2537.

[3] Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2012, 20(1): 30–42. [doi: 10.1109/TASL.2011.2134090]

[4] Gu TY, Liu K, Dolan-Gavitt B, Garg S. BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019, 7: 47230–47244. [doi: 10.1109/ACCESS.2019.2909068]

[5] Liu YF, Ma XJ, Bailey J, Lu F. Reflection backdoor: A natural backdoor attack on deep neural networks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 182–199. [doi: 10.1007/978-3-030-58607-2_11]

[6] Li YM, Zhai TQ, Jiang Y, Li ZF, Xia ST. Backdoor attack in the physical world. arXiv:2104.02361, 2021.

[7] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 2206–2216.

[8] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[9] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: ICLR, 2018.

[10] Bai JW, Chen B, Li YM, Wu DX, Guo WW, Xia ST, Yang EH. Targeted attack for deep hashing based retrieval. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 318–364. [doi: 10.1007/978-3-030-58452-8_36]

[11] Wu DX, Xia ST, Wang YS. Adversarial weight perturbation helps robust generalization. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 249.

[12] Bai Y, Zeng YY, Jiang Y, Xia ST, Ma XJ, Wang YS. Improving adversarial robustness via channel-wise activation suppressing. In: Proc. of the 9th Int'l Conf. on Learning Representations. ICLR, 2021.

[13] Li YM, Wu BY, Feng Y, Fan YB, Jiang Y, Li ZF, Xia ST. Semi-supervised robust training with generalized perturbed neighborhood. Pattern Recognition, 2022, 124: 108472. [doi: 10.1016/j.patcog.2021.108472]

[14] Zhe Z, Tang D, Wang XF, Han WL, Liu XY, Zhang KH. Invisible mask: Practical attacks on face recognition with infrared. arXiv:1803.04683, 2018.

[15] Zhang YY, Deng WH. Towards transferable adversarial attack against deep face recognition. IEEE Trans. on Information Forensics and Security, 2020, 16: 1452–1466. [doi: 10.1109/TIFS.2020.3036801]

[16] Barni M, Kallas K, Tondi B. A new backdoor attack in CNNS by training set corruption without label poisoning. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing. Taipei: IEEE, 2019. 101–105. [doi: 10.1109/ICIP.2019.8802997]

[17] Tang RX, Du MN, Liu NH, Yang F, Hu X. An embarrassingly simple approach for Trojan attack in deep neural networks. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 218–228. [doi: 10.1145/3394486.3403064]

[18] Li SF, Xue MH, Zhao BZH, Zhu HJ, Zhang XP. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Trans. on Dependable and Secure Computing, 2020, 18(5): 2088–2105. [doi: 10.1109/TDSC.2020.3021407]

[19] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks. arXiv:1912.02771, 2019.

[20] Nguyen TA, Tran AT. WaNet-imperceptible warping-based backdoor attack. In: Proc. of the 9th Int'l Conf. on Learning Representations. ICLR, 2021.

[21] Quiring E, Rieck K. Backdooring and poisoning neural networks with image-scaling attacks. In: Proc. of the 2020 IEEE Security and Privacy Workshop. San Francisco: IEEE, 2020. 41–47. [doi: 10.1109/SPW50608.2020.00024]

[22] Xiao QX, Chen YF, Shen C, Chen Y, Li K. Seeing is not believing: Camouflage attacks on image scaling algorithms. In: Proc. of the 28th USENIX Security Symp. Santa Clara: USENIX, 2019. 443–460.

[23] Wenger E, Passananti J, Bhagoji AN, Yao YS, Zheng HT, Zhao BY. Backdoor attacks against deep learning systems in the physical world. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3202–3211. [doi: 10.1109/CVPR46437.2021.00614]

[24] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 1505–1521.

[25] Shumailov I, Shumaylov Z, Kazhdan D, Zhao YR, Papernot N, Erdogdu MA, Anderson RJ. Manipulating SGD with data ordering attacks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. NeurIPS, 2021. 18021–18032.

[26] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pretrained models. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 2793–2806. [doi: 10.18653/v1/2020.acl-main.249]

[27] Dong YP, Yang X, Deng ZJ, Pang TY, Xiao ZH, Su H, Zhu J. Black-box detection of backdoor attacks with limited information and data. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 16462–16471. [doi: 10.1109/ICCV48922.2021.01617]

[28] Chou E, Tramèr F, Pellegrino G. SentiNet: Detecting localized universal attacks against deep learning systems. In: Proc. of the 2020 IEEE Security and Privacy Workshop. San Francisco: IEEE, 2020. 48–54. [doi: 10.1109/SPW50608.2020.00025]

[29] Chen HL, Fu C, Zhao JS, Koushanfar F. Deepinspect: A black-box Trojan detection and mitigation framework for deep neural networks. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI, 2019. 4658–4664. [doi: 10.24963/ijcai.2019/647]

[30] Shen GY, Liu YQ, Tao GH, An SW, Xu QL, Cheng SY, Ma SQ, Zhang XY. Backdoor scanning for deep neural networks through K-arm optimization. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 9525–9536.

[31] Liu YT, Xie Y, Srivastava A. Neural Trojans. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Design. Boston: IEEE, 2017. 45–48. [doi: 10.1109/ICCD.2017.16]

[32] Udeshi S, Peng SS, Woo G, Loh L, Rawshan L, Chattopadhyay S. Model agnostic defence against backdoor attacks in machine learning. IEEE Trans. on Reliability, 2022, 72(2): 880–895. [doi: 10.1109/TR.2022.3159784]

[33] Villarreal-Vasquez M, Bhargava B. ConFoc: Content-focus protection against Trojan attacks on neural networks. arXiv:2007.00711, 2020.

[34] Yoshida K, Fujino T. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In: Proc. of the 13th ACM Workshop on Artificial Intelligence and Security. ACM, 2020. 117–127. [doi: 10.1145/3411508.3421375]

[35] Chen XY, Liu C, Li B, Lu K, Song D. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526, 2017.

[36] Bagdasaryan E, Veit A, Hua YQ, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2018. 2938–2948.

[37] Chen CL, Golubchik L, Paolieri M. Backdoor attacks on federated meta-learning. arXiv:2006.07026, 2020.

[38] Zhu C, Huang WR, Li HD, Taylor G, Studer C, Goldstein T. Transferable clean-label poisoning attacks on deep neural nets. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7614–7623.

[39] Souri H, Fowl L, Chellappa R, Goldblum M, Goldstein T. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: NIPS, 2021.

[40] Liu RS, Gao JX, Zhang J, Meng DY, Lin ZC. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10045–10067. [doi: 10.1109/TPAMI.2021.3132674]

[41] Garg S, Kumar A, Goel V, Liang YY. Can adversarial weight perturbations inject neural backdoors. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. ACM, 2020. 2029–2032. [doi: 10.1145/3340531.3412130]

[42] Zhang J, Chen DD, Liao J, Huang QD, Hua G, Zhang WM, Yu NH. Poison ink: Robust and invisible backdoor attack. arXiv: 2108.02488, 2021.

[43] Li YZ, Li YM, Wu BY, Li LK, He R, Lyu SW. Invisible backdoor attack with sample-specific triggers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 16443–16452. [doi: 10.1109/ICCV48922.2021.01615]

[44] Yao YS, Li HY, Zheng HT, Zhao BY. Regula sub-rosa: Latent backdoor attacks on deep neural networks. arXiv:1905.10447, 2019.

[45] Tan TJL, Shokri R. Bypassing backdoor detection algorithms in deep learning. In: Proc. of the 2020 IEEE European Symp. on Security and Privacy (EuroS&P). Genoa: IEEE, 2020. 175–183. [doi: 10.1109/EuroSP48549.2020.00019]

[46] Zhong HT, Liao C, Squicciarini AC, Zhu SC, Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation. In: Proc. of the 10th ACM Conf. on Data and Application Security and Privacy. New Orleans: ACM, 2020. 97–108. [doi: 10.1145/3374664.3375751]

[47] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 86–94. [doi: 10.1109/CVPR.2017.17]

[48] Doan K, Lao YJ, Zhao WJ, Li P. LIRA: Learnable, imperceptible and robust backdoor attacks. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 11946–11956. [doi: 10.1109/ICCV48922.2021.01175]

[49] Doan K, Lao YJ, Li P. Backdoor attack with imperceptible input and latent modification. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. NeurIPS, 2021. 18944–18957.

[50] Koren Y. Factor in the neighbors: Scalable and accurate collaborative filtering. ACM Trans. on Knowledge Discovery from Data, 2010, 4(1): 1. [doi: 10.1145/1644873.1644874]

[51] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4949–4958. [doi: 10.1109/CVPR.2019.00509]

[52] Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. In: Proc. of the 37th Asilomar Conf. on Signals, Systems & Computers, 2003. Pacific Grove: IEEE, 2003. 1398–1402. [doi: 10.1109/ACSSC.2003.1292216]

[53] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: 10.1109/CVPR.2018.00068]

[54] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

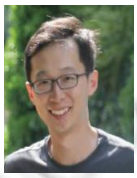
[55] Christian S, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 4278–4284.

[56] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. ICLR, 2021.

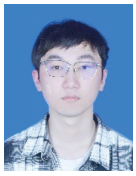
吴尚锡(1997－), 男, 博士生, 主要研究领域为可信赖机器学习.

陈观浩(2001－), 男, 本科生, 主要研究领域为机器学习, 计算机视觉.

尹雨阳(2000－), 男, 本科生, 主要研究领域为机器学习, 计算机视觉.



桑基韬(1985－) 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为可信赖机器学习, 人工智能及应用.



宋思清(2001－), 男, 本科生, 主要研究领域为机器学习, 计算机视觉.

于剑(1969－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为机器学习与认知计算, 人工智能及应用.