

基于 RSA 的图像可识别对抗攻击方法

张宇, 李海良

(暨南大学, 广东 广州 510632)

摘要: 基于密码学中的 RSA 签名方案与 RSA 加密方案, 提出了一种能够让特定分类器输出对抗样本正确分类的对抗攻击方法。通过单像素攻击的思想使正常图像在嵌入附加信息的同时能够具有让其余分类器发生错误分类的能力。所提方法可以应用在分类器授权管理与在线图像防伪等领域。实验结果表明, 所提方法生成的对抗样本对于人眼难以察觉, 并能被特定分类器识别。

关键词: 对抗攻击; 签名方案; 加密方案; 密码学; 分类器; 图像识别

中图分类号: TP181

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2021065

RSA-based image recognizable adversarial attack method

ZHANG Yu, LI Hailiang

Jinan University, Guangzhou 510632, China

Abstract: Adversarial attack is an important part of deep learning security research. Relying on the RSA signature schemes and RSA encryption schemes in cryptography, an adversarial attack method that adversarial examples can be recognized by a specific classifier is proposed. Through the idea of one pixel attack, the normal image can have the ability to make other classifier misclassify while embedding additional information. It can be used in classifier authorization management, online image anti-counterfeiting, etc. The experiment show that the adversarial examples can be recognized under the specific classifier, and the disturbance noise is difficult to detect by the human eye.

Keywords: adversarial attack, signature scheme, encryption scheme, cryptography, classifier, image recognition

1 引言

随着深度学习技术^[1]的研究与发展, 越来越多的领域适应了深度学习技术带来的巨大变革, 各类深度学习应用如图像分类^[2]、人脸识别^[3]、语

音识别^[4]、自然语言处理^[5]等进入了人们的生活, 为人们的工作和生活提供便利。这些技术的广泛应用带动了人工智能的普及与发展, 但同时这些应用面临的一些安全性问题引起了研究者的关注与探讨, 这些安全性问题主要涉及数据隐私与错

收稿日期: 2020-11-01; 修回日期: 2021-03-15

通信作者: 李海良, lihailiang@jnu.edu.cn

基金项目: 广东省重点研发计划 (2020B0101090004); 广东省科技创新战略专项资金 (pdjh2021b0058)

Foundation Items: Key Research and Development Program of Guangdong Province (2020B0101090004), Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation (pdjh2021b0058)

论文引用格式: 张宇, 李海良. 基于 RSA 的图像可识别对抗攻击方法[J]. 网络与信息安全学报, 2021, 7(5): 40-48.

ZHANG Y, LI H L. RSA-based image recognizable adversarial attack method[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 40-48.

误分类等方面。对这些安全性问题的研究是深度学习应用能够进一步推广落实的重要前提, 一个良好的深度学习应用需要使用大量的数据进行训练, 训练过程所需要的时间成本与硬件成本是使用者必须考虑的问题。训练完成后得到的各项参数以及使用的深度学习网络模型结构往往成为商业应用中的机密信息。使用者不希望花费了大量资源的深度学习应用面临隐私泄露或在实际使用中发生分类错误的风险, 因为这些风险意味着巨大的信誉下降和财产损失。

在深度学习的安全性研究中, 对抗攻击是非常重要的研究方向, 也是研究的热点之一。对抗攻击和对抗样本的概念最早由 Szegedy 在 2013 年提出: 一个经过良好训练的深度学习分类器会以高置信度输出一幅看起来十分正常的图片的错误分类, 这类图片被命名为对抗样本^[6]。

图 1 为一张 Imagenet 数据集^[7]的正常图像, 以及对其使用快速梯度下降法^[6] (fast gradient sign method) 生成的对抗样本, 其原图像的识别结果是“鸵鸟”。但在同一个分类器下, 其添加了噪声后所产生的对抗样本识别结果却是“起重机”。这类扰动噪声尽管人眼难以察觉, 但对于计算机来说却非常明显。如图 1 所示, 人眼中相差无几的图像对于计算机而言差别极大, 以至于可使一个经过大量数据训练的分类器做出错误的判断。如果放在各类商业应用中, 当人脸识别、自动驾驶等过程发生识别错误, 则必然会造成巨大的损失, 可见这类人眼不可察觉的扰动噪声对深度学习分类器的威胁是巨大的。



图 1 对抗样本及其噪声
Figure 1 Adversarial example and noise

在 2019 年的一份研究机器学习安全性的综述文章中, 作者所调查的论文中有关对抗攻击的文章占总数的 50%^[8], 这说明研究者同样认可对抗攻击对于机器学习安全性的巨大威胁。但目前

对抗攻击的研究^[9-13]大部分集中在如何更高效地生成对抗样本; 如何生成隐蔽性更小的对抗样本; 如何生成具有物理鲁棒性的对抗样本等。而对于对抗样本的识别和防御更多集中在通过针对性训练^[14-15]、过滤^[16]、蒸馏^[17]、压缩还原^[18]等方法, 来提高分类器的鲁棒性。

这些攻击方法与防御策略的出发点都是希望减少对抗攻击对深度学习分类器的威胁。在研究了多种对抗攻击算法与防御策略后, 本文提出了一种能够令特定分类器输出对抗样本正确分类的对抗攻击方法。该方法结合了密码学中的签名与加密机制, 能够保证具有密钥的深度学习分类器正确输出由本文方法生成的对抗样本正确分类。而在不具有密钥的分类器中, 这些对抗样本则会让该分类器发生分类错误。

结合以上特性, 本文提出的对抗攻击方法有望能应用在分类器授权管理与图像防伪等领域。在数据资产越来越重要的今天, 数据私有化与信息安全成为商业应用中必须考虑的问题。本文提出的对抗攻击能够使获得密钥的分类器拥有对本类对抗样本完全的防御能力, 而密钥的分配由使用者决定, 可以结合其余密码学方法构建对多个分类器的授权机制。获得授权的分类器能够得到对应的密钥, 用以对输入的图像进行识别, 在密钥不泄露的情况下能够防止非授权用户通过窃取深度学习模型参数和网络结构来获取利益, 如图 2 所示。

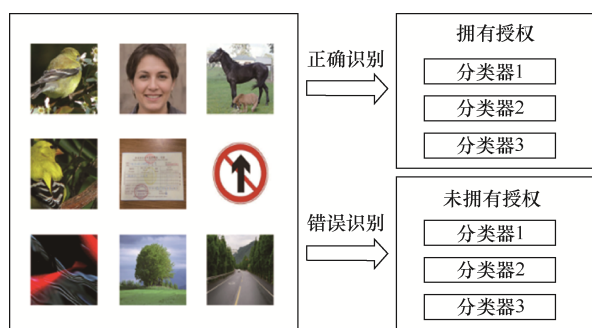


图 2 分类器授权管理
Figure 2 Classifier authorization management

本文方法的主要贡献和创新可以总结为 3 个方面。

- ① 将密码学与对抗攻击相结合, 提出了一种基于密码学困难性问题的可特定识别对抗攻击方法。
- ② 现有的对抗攻击防御策略中效果最佳的

针对性训练难以达到对对应类别对抗攻击的完全防御,会有较多数量的对抗样本成功地让分类器发生错误分类,但本文方法产生的对抗样本能够被特定的分类器识别。

③ 本文方法中使用到的签名信息和密文信息受到实际使用的密码学方案的影响,其长度和形式均有较强的可拓展性。在未来的研究中可以通过使用不同的密码学方案来实现更好的攻击效果或提高生成效率。

2 相关工作

本节主要介绍所提对抗攻击方法涉及的相关定义与研究现状。

2.1 相关定义介绍

(1) 对抗攻击

对抗攻击是一种攻击深度学习模型的攻击方法^[19],可由下述方式定义。

对于一张正常图像 (x, y) , 其中 x 为正常图像, y 为其对应的分类。攻击者通过构造:

$$x' \leftarrow x + \theta \quad (1)$$

得到一张添加了噪声的图像,其中 θ 为噪声。 θ 的构造和范围由具体的对抗攻击方法决定。攻击者所输出的对抗样本 (x', y') 应满足

$$y' \neq y \quad (2)$$

根据攻击者对目标模型信息的掌握程度,对抗攻击可以分为三类^[20-22]: 黑盒攻击、灰盒攻击和白盒攻击。

黑盒攻击: 攻击者不了解目标模型的结构和参数,只能通过正常的输入得到模型的输出。通过不断地与模型进行正常交互来设计与实现攻击。

灰盒攻击: 攻击者能够得知目标模型的结构和参数,并在这些信息的基础上构造攻击算法和实现攻击。

白盒攻击: 攻击者能够得知目标模型的所有信息,包括但不限于模型的结构、参数、训练集、测试集等信息。通常来讲,白盒攻击对模型的攻击效果最好。

根据对抗攻击的攻击结果,对抗攻击可以分为定向攻击和非定向攻击。前者的目的是让目标深度学习分类器输出攻击者指定的分类结果;后

者的目的是让目标深度学习分类器输出错误的分类结果。

(2) 签名机制与加密机制

一个数字签名方案通常由 3 个算法组成^[23]: 密钥生成算法、签名算法和确定性验证算法。

密钥生成算法 Gen 负责输出一对密钥 (vk, sik) , vk 称为验证密钥, sik 称为签名密钥。

签名算法 Sign 接收签名密钥 sik 与消息 m 作为输入, 输出一个签名 σ 。

确定性验证算法 Vrfy 接收验证密钥 vk 、消息 m 与签名 σ 作为输入, 输出验证结果。如果签名有效则验证通过, 输出 1; 签名无效则验证失败, 输出 0。

公钥加密方案的思想最早由 Diffie 和 Hellman 提出^[24], 最早实现是由 Rivest、Shamir 和 Adleman 在 1978 年提出的 RSA 方法^[25]。

常用的公钥加密方案通常由 3 个算法构成: 密钥生成算法、加密算法和解密算法。

密钥生成算法 Gen 输出一个密钥对 (pk, sk) , pk 称为公钥, sk 称为私钥。

加密算法 Enc 接收公钥 pk 与加密的明文 m 作为输入, 输出一个密文 c 。

解密算法 Dec 接收私钥 sk 与解密的密文 c 作为输入, 输出一个明文 m 。

使用公钥加密方案能够保证明文信息不被敌手获知, 敌手唯一能够获得明文信息的方式就是解密。但在不具有私钥的情况下, 一个可证明安全性的公钥加密方案是难以被攻破的。在本文方法中对抗样本真实分类的标签被加密为密文, 以此来避免信息泄露。

2.2 研究现状

对抗攻击能够直接影响深度学习分类器的识别能力, 成为深度学习安全研究的热点, 目前, 对于对抗攻击的研究分为攻击与防御两方面。

在对抗攻击的攻击方面, 主要的对抗攻击方法有基于梯度的攻击、基于分数的攻击与基于决策的攻击。基于梯度的攻击的主要思路是在白盒环境下, 对输入的图像像素数据进行求导, 得到具体的梯度方向, 根据该方向进行扰动噪声的计算, 使生成的对抗样本发生错误分类; 基于分数的攻击与基于决策的攻击主要应用在黑盒环境

下,通过多次查询模型得到的信息进行计算。

在防御对抗攻击方面,目前较为良好的防御方式是针对性训练,该方式主要在训练阶段加入对抗样本数据,通过训练过程让模型学习对抗样本与正常图片间的细微差别,从而达到在使用阶段可以正确识别该类对抗样本的目的。2020 年的一篇探究图像识别分类器鲁棒性的论文指出^[26]:目前在 *cafir10* 数据集上较好的针对性训练方法^[27]能达到的识别精度为 87.3%,在 *Imagenet* 数据集上效果较好的方法能达到的精度则为 73.5%^[14]。可以看到,仍然难以达到 100% 的识别精度。

2018 年,在识别特定的对抗样本方面,Hyun 等^[28]提出了一种多目标对抗攻击方法,产生的对抗样本能够被友善的分类器正确分类,其应用场景主要针对军事战略领域。作者指出对抗攻击可以用来欺骗敌方的分类器,如修改战场上的路标以欺骗敌人的自动驾驶汽车。但该对抗样本的产生主要通过机器学习方法来实现,并且对于图像,计算出敌方分类器与友方分类器之间的识别差异,需要较长的运行时间。同时,在计算过程中需要了解模型的内部信息,但在商业行为中的深度学习模型信息往往被视为商业机密而加以保护,使用者通常难以得到其模型的内部信息。

3 对抗攻击方法

本文所提的对抗攻击方法可以概括为:将一张能够被分类器正确识别的图像通过嵌入附加信息的方式改造为只有在特定分类器中才能正确识别的对抗样本。与一般的对抗攻击方法相比,本文方法中的扰动噪声包含了原图像的两项重要信息(图像签名信息与正确分类信息),这两项信息通过密码学中的签名方案和加密方案生成,在处理后通过隐蔽的方式嵌入原图像,如图 3 所示。

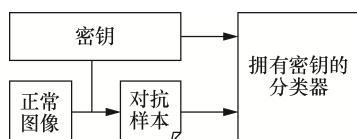


图3 本文提出的对抗攻击的流程
Figure 3 The flow of our adversarial attack

本文方法是一种非定向黑盒攻击,不指定攻击的目标分类,同时攻击过程不涉及目标深度学

习模型的内部结构和参数,仅通过反复与目标模型进行正常交互来获得攻击所需的信息。

3.1 RSA 方案

本文方法结合了密码学中的签名方案与加密方案,在目前的实例化中采用的具体算法是 RSA 方法。

RSA 的签名方案与加密方案,基本形式与一般的数字签名方案和公钥加密方案一致,其安全性以大整数分解的困难性为前提。具体的方案构造如下。

算法 GenRSA 用来生成基于 RSA 问题的密钥参数。该算法接收一个安全参数,输出一组参数 $(N, e, d) \in \mathbb{N}^3$, 同时参数 N, e, d 满足

$$e \cdot d \bmod \varphi(N) = 1 \quad (3)$$

其中, $\mathbb{N}$ 为非负整数集合, $\varphi(\cdot)$ 是欧拉函数。

基于 GenRSA 算法, RSA 的数字签名方案如下。

密钥生成算法 Gen 接收一个安全参数作为输入,将安全参数输入 GenRSA 中,得到一组参数 (N, e, d) 。令验证密钥 vk 为 (N, e) , 签名密钥 sik 为 (N, d) 。

签名算法 Sign 接收签名密钥 sik 和消息 m 作为输入,计算签名 $\sigma := [H(m)]^d \bmod N$ 。

验证算法 Vrfy 接收验证密钥 vk , 消息 m 与签名 σ 作为输入,在 $\sigma^e \bmod N = H(m)$ 时输出 1, 其余情况输出 0。

$H(\cdot)$ 为摘要函数,接收任意长度的比特串输入,输出一个固定长度的 hash 函数值。

基于 RSA 的公钥加密方案如下。

密钥生成算法 Gen 接收一个安全参数作为输入,将安全参数输入 GenRSA 中,得到一组参数 (N, e, d) 。令公钥 pk 为 (N, e) , 私钥 sk 为 (N, d) 。

加密算法 Enc 接收公钥 pk 和消息 m 作为输入,计算密文 $c := m^e \bmod N$ 。

解密算法 Dec 接受私钥 sk 与密文 c 作为输入,计算明文 $m := c^d \bmod N$ 。

3.2 对抗攻击方案

基于 RSA 方案,本文对抗样本的生成与识别可划分为以下 4 个阶段。

(1) 密钥生成阶段

在这一阶段通过密钥生成算法 Gen, 输入安全参数 λ , 输出一组密钥对 $(key_{vrfy\&enc}, key_{sign\&dec})$,

其中, $\text{key}_{\text{vrfy}\&\text{enc}}$ 对应签名方案中的验证密钥与公钥加密方案中的加密密钥, $\text{key}_{\text{sign}\&\text{dec}}$ 对应签名机制中的签名密钥与公钥加密机制中的解密密钥。这两个密钥是一组密钥对, 需要使用者秘密保有, 并通过安全的渠道传输给相应的分类器, 该分类器正是上文提到的“拥有授权的分类器”。

分类器在得到密钥后应该将其存储在安全的环境中, 因为分类器中密钥泄露导致的安全性问题暂不在本文方法的考虑范围内。

(2) 图像信息获取阶段

这一阶段的目的是获得输入的正常图像的签名信息与正确分类信息。本文方法使用 RSA 签名算法与 RSA 加密算法进行签名与加密, 并使用 MD5 算法对图像进行 hash 运算, 得到可以进行 RSA 签名的图像 hash 函数值。

假设输入图像的像素矩阵为 Arr_{img} , 其真实分类为 y 。为了保证签名算法中签名对象的不变性, 此处仅对图像首行和尾行以外的像素进行签名, 被签名的像素称为图像的主体像素, 图像的首行像素和尾行像素将作为签名信息和分类信息的嵌入位置。

算法过程如下。

算法 1 图像签名信息与分类信息提取方法

1) 获得主体像素的信息 hash 函数值 $h \leftarrow \text{MD5}(\text{Arr}_{\text{img}}[1:-1])$ 。

2) 对 hash 函数值 h 进行签名, 得到签名 $\partial \leftarrow \text{Key}_{\text{sign}\&\text{dec}} h$ 。

3) 对正确分类 y 进行加密, 得到密文 $c \leftarrow \text{key}_{\text{vrfy}\&\text{enc}} y$ 。

4) 输出图像主体签名 ∂ 与正确分类密文 c 。

原图像信息提取过程如图 4 所示。

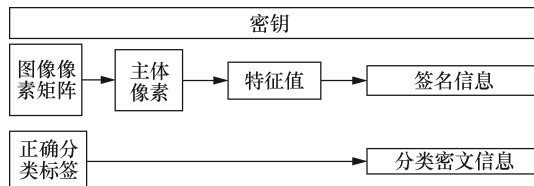


图 4 原图像信息提取过程
Figure 4 Clean image information extraction process

关于本阶段涉及的签名与密文长度, 以 224×224 的图像举例, 其行长为 224。目前的实例化采用的 RSA 算法产生的密钥长度为 1 024 bit,

签名与密文长度均为 172 bit。

(3) 扰动像素嵌入阶段

这一阶段的目的是得到足以让分类器发生错误分类的扰动, 并将扰动像素和上一阶段得到的签名信息和分类密文信息一同嵌入图像的首行和尾行。签名信息与正确分类信息分别和扰动像素组成向量, 称为签名扰动向量与分类扰动向量。签名扰动向量替换图像首行像素的第一通道; 分类扰动向量替换图像尾行像素的第一通道。

在目前的实例化中, 上一阶段得到的签名信息和正确分类密文均为一串长度固定的字符串, 字符串中的字符会转化为 ASCII 码, 与计算得到的扰动像素一同组成长度为图像行向量长度的向量。一条扰动向量的基本构成如下:

$$\text{disturbance vector} \leftarrow \text{inf} + \text{noisy} + \text{oripixel} \quad (4)$$

其中, inf 为上一阶段得到的签名与密文信息对应的 ASCII 码向量, noisy 为本阶段反复迭代测试得到的足以让分类器发生错误分类的像素向量, oripixel 为原图像行向量中不改动的位置的像素向量。

扰动向量生成过程如图 5 所示。

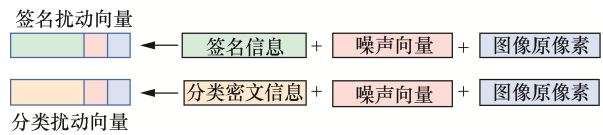


图 5 扰动向量生成过程
Figure 5 Disturbance vector generation process

可以看到, 签名信息和分类密文信息对应的像素起始位置为得到的扰动向量的起始位置, 噪声向量与图像原像素则直接连在签名信息与分类密文信息之后。

以 224×224 的图像为例, 签名信息与分类密文信息均为 172 bit, 则两个扰动向量中的前 172 bit 分别用来存放两种图像附加信息, 之后的 52 bit 则用来存放噪声向量和图像原像素。

本文方法中噪声像素的生成方法借鉴了单像素攻击^[30]的攻击思路。这类对抗攻击的主要思想是在黑盒模型下反复改动图像中的像素, 每改动一次就查询一次分类器查询分类, 直到找到能够让分类器发生错误分类的像素为止。

在本文方法的理想实现中, 算法会将图像首行与尾行像素中除去包含签名像素与分类信

息像素的剩余部分视为噪声生成位置,在保持其余像素不变的情况,按照从小到大的位置关系对图像像素进行改动,其最终输出的像素值大小范围为 0~255。但为了算法的运行效率,本文在实例化中统一规定噪声向量的长度为 32 bit。每次迭代算法都会生成一串长度为 32 bit 的随机像素向量,通过扩大噪声位置来提高噪声像素的攻击能力。假如直到迭代上限也没有找到能够让图像发生错误分类的噪声,则视为攻击失败。

从理论的角度分析,多目标分类的深度学习模型大部分的分类平面极其复杂,当对输入的实例做出细微差别后,其与分类边界的距离必然会发生改动,所以通过反复迭代修改固定位置的单个像素或多个像素均有相当大的概率让该实例越过周围的某个分类边界,从而发生分类变化。这一点已经在实验中得到验证,本文方法正是通过反复迭代修改首行像素与尾行像素中不涉及签名信息与分类信息的像素,直到找到能够让分类器识别错误的像素扰动信息。

最后,本阶段输出一张替换了首行与尾行像素的图像作为本文方法生成的对抗样本。

(4) 识别阶段

这一阶段的目的是验证生成的对抗样本能够被拥有密钥的分类器正确验证,并输出正确的分类信息。

具有密钥的分类器在接收到一张图像的输入后,会先将该图像分成三部分,分别对应图像首行信息、图像主体信息与图像尾行信息。假如该图像是本文方法生成的对抗样本,因为签名信息与密文信息存放的位置固定,且均在第一通道,则分类器可以简单地从图像首行与尾行的第一通道中得到签名信息和分类信息,并使用存储的密钥对签名信息和图像主体像素信息进行验证,验证通过后再使用同一密钥对的另一密钥对分类信息进行解密,得到图像的真实分类结果并直接输出;假如输入的图像是一张正常图像,则分类器同样进行上述步骤,但在签名验证阶段无法得到签名通过的结果,分类器此时会直接对图像进行深度学习识别过程,输出图像的识别分类。

4 安全性分析

本文提出的可识别对抗攻击的核心是密码学中的签名与加密方案,其理想情况是能够对一张正常图像输出其可识别的对抗样本。拥有密钥的分类器能够识别这一对抗样本,同时这种对抗攻击只能由拥有密钥的使用者来实行。

本文对抗攻击的构造可以让敌手难以伪造出这类对抗样本。为了定义和分析其安全性,假设存在一个敌手希望伪造一张对抗样本来欺骗拥有密钥的分类器,则一个简单的安全性实验序列设计如下。

算法 2 对抗样本伪造攻击

1) 挑战者运行密钥生成算法 Gen 得到一组密钥对 $(\text{key}_{\text{vrfy\&enc}}, \text{key}_{\text{sign\&dec}})$, 对于输入的正常图像,使用 $\text{key}_{\text{sign\&dec}}$ 对其主体像素进行签名,得到签名 σ , 使用 $\text{key}_{\text{vrfy\&enc}}$ 对其分类进行加密,得到密文 c 。计算出签名扰动向量与分类扰动向量后将其嵌入原图像的首行与尾行,得到对抗样本 adv 。

2) 敌手 A 获得该对抗样本 adv , A 可以通过挑战者向拥有密钥的分类器询问图像能否通过验证和分类器对图像的输出分类。记 Q 为 A 所询问过的所有图像的集合。

3) A 通过已有的信息输出一张新的图像 adv' 。

4) 如果 $\text{adv}' \notin Q$, 并且可以通过分类器的验证并输出符合分类格式的结果,则认为 A 获胜,输出 1; 其余情况输出 0。

如果对于任意敌手 A , 均有

$$\Pr[\text{Expt}(\lambda) = 1] \leq \text{negl}(\lambda) \quad (5)$$

则认为对于本文方法产生的对抗样本,敌手难以伪造出另一张对抗样本欺骗拥有密钥的分类器。其中, λ 是安全参数,影响密钥、密文和签名的长度。 $\text{negl}(\cdot)$ 为一个可忽略函数。这条不等式代表 A 赢得游戏的概率可以忽略。

当敌手希望伪造出一张对抗样本欺骗拥有密钥的分类器时,其必然需要构造出能够通过验证的签名向量和能够正确解密出分类信息的正确分类向量。本文方法目前的实例化采用 RSA 算法,敌手必须要克服 RSA 困难问题,才能够构造出顺

利通过验证和解密的图像。目前,很难找到一个拥有足够计算能力的敌手来实现对本方法中生成的对抗样本进行伪造。

如果敌手期望通过计算以外的方式来生成对抗样本,则可能选用的方法是利用其得到的对抗样本中的签名信息和密文信息。因为签名向量和分类向量被直接替换为图像首行与尾行的第一通道,敌手可以直接通过替换像素的方式将已有的签名向量和分类向量替换到其构造的图像上。但这种替换像素的伪造方法依然无法绕过 RSA 困难性问题,因为签名验证过程需要首行像素信息与主体像素信息共同参与,同时密文解密过程使用的密钥必然与签名过程使用的验证密钥相匹配。敌手只有在同时替换了首行像素和主体像素的情况下才能通过验证过程,但分类密文解密过程仍然需要面对 RSA 困难问题。在不具备量子计算能力与密钥的前提下,敌手所伪造的密文会以极大概率解密得到一串无意义的字符串或无法解密,依然会被分类器察觉。而其余像素替换方法均无法绕开验证过程的 RSA 困难问题。

5 实验

本文方法在实例化中使用的是 RSA 签名方案与 RSA 加密方案,其使用的密钥长度为 1 024 bit,输出的签名和密文长度均为 172 bit。

实验的软硬件环境为 4 GPU 的 Linux 服务器,系统为 64 位 Ubuntu18.04.4LTS,内存 256 GB,处理器为 Intel Xeon CPU E5-2683 V3 @ 2.00 GHz, GPU 型号是 Gtx 1080ti,显存 12 GB。使用的数据集为 Imagenet 数据集,实验中的图像尺寸均为 224×224。使用的 RSA 签名方案代码与 RSA 加密方案代码由 Python 中的 Crypto 库提供。

为了衡量本文方法的攻击成功率,本文对随机 1 000 张 ImageNet 的图像进行了模拟攻击实验。实验使用的模型为 Torchvision 提供的 VGG-16 预训练模型。在对图像进行识别前,图像的首尾两行像素被替换为噪声像素。随着迭代次数的增加,模型的识别结果发生了改变,其中,原始识别标签的图像数量变化如表 1 所示。

从表 1 中可以看到,迭代次数的增加能提高模型攻击的成功率,但由于噪声位置的限制,一张可识别对抗样本的生成通常需要较高的迭代次数。

表 1 迭代次数对攻击成功率的影响

Table 1 The impact of iteration times on attack success rate

迭代次数	攻击成功数	攻击成功率
0	370	37%
5	707	70.7%
10	771	77.1%
15	799	79.9%
20	808	80.8%

实验的实例化输出如下:实验中生成的对抗样本均可在拥有密钥的分类器中被正确验证并通过分类密文解密得到对应的正确分类。

图 6 为本文方法中的一个输出实例,图内正常图片的分类标签为 13,生成的对抗样本的分类标签为 16。

正常图片	对抗样本	签名信息	分类密文
		If38ipNffu3oqkzX01 SjTw3EYkxRy11OL Ky4oPZtwg0hVkt55 ufi7+5JlTb+/idZ6C1 NCxrr+jANktmPT+ Pywz8rVszASnjT3k o17kRtVq+SkNOp0 HEhS8M/sXb5sSZg 1LLJsU VdsVDevev 8MDckf6GBec5rp5 XGp8mtAThDlhDI vU=	fSf4JKN10nH/kPdi0 w4a4SBjc+objJTDa V25SsekjeFCF5GR JrAjttiH+xqt5t/TPcP NsVtB3Bgscqvvt0h LrrgJInCEQqWLEQ ysFbQqMZYBQax4 RsJIf5Sx7bhgVbO Hs6SOWuSoaRdv cv tju7TivxChSvTiOO+ +YLZFi3zixp4=
md5特征码 d798c65b237f6e34d48ef 2d1f5c94f84			

图 6 输出实例 1

Figure 6 Output example 1

图 7 为另一个输出实例,图内正常图片的分类标签为 13,生成的对抗样本的分类标签为 20。

正常图片	对抗样本	签名信息	分类密文
		WQ6JZOvpA0/zB6d vbZg8ER38Tt0DYs NVmLT1ppKFatLC 8cgKWwpz01uccXg UekDsGHmSp/Mhq WfkWzuok558oohW CzlkTYQfs1LVjgRm VFsTs9iAohBtHw1D Fi/CWmwUZFnWm1 9tpsr/HoXvcsIU+bg AZ4cZ7GxjeplpvMQ ICs=	YBlzfN6XG9R7/mg KM114sjn1Be5XTp kV8FFULn4C35815 zOM/vCqGd2om0x T8QwQ78MRwQW Gap+ZDQVrMTpqn SfW/uoQN1Ytz8pwQ 7mxxqa+8wuzEYN6 Kc8yN1ZG+9n3hbx IOxj2lgwCiThbYlR Nq2xVdaqNIGdpH GRPMx4F6PRU=
md5特征码 92a414d7cece6798bcf964 ad6fd5941e			

图 7 输出实例 2

Figure 7 Output example 2

输出实例中的签名信息与分类密文可以被正常读取，敌手在无法伪造能顺利通过验证的签名与能够解密的密文的情况下难以伪造本对抗攻击。

图 8 为测试中输出其余部分图片的效果展示，图内的数字为该图像的识别分类结果。可以看出，尽管本文方法相比单像素攻击改动的像素较多，但依然具备人眼难以察觉的特性，同时能让分类器发生错误分类。



图 8 输出实例 3
Figure 8 Output example 3

本文方法与其他 4 种经典对抗攻击方法的简单对比总结如表 2 所示。

表 2 对抗攻击方法对比
Table 2 Comparison of several adversarial attacks method

对抗攻击	噪声	可特定识别性
本文方法	408~448 bit	√
单像素攻击	1bit	×
快速梯度下降攻击	全局像素	×
Deepfool 攻击	全局像素	×
C&W 攻击	全局像素	×

其中，Deepfool 攻击生成的扰动是模型的最小近似扰动，其扰动噪声的范围针对全局像素；C&W 攻击的全称是 Carlini&Wagner 攻击，是一种迭代优化的低扰动对抗样本生成算法，其扰动噪声发生的范围同样是全局像素。从扰动的角度来分析，除单像素攻击与本文方法以外的 3 种经典攻击方法的目的是生成足以让分类器发生错误分类的扰动，这需要对图像上的所有像素进行计算和分析，得到期望的最优扰动噪声并分布在图像的所有像素中。而单像素攻击与本文方法则对扰动噪声生成的位置和范围做出了限制，单像素

攻击要求生成的扰动噪声只存在一个像素上，本文方法生成的扰动噪声只存在两行像素向量中，从像素改动个数的角度来看，本文对抗攻击方法改动的像素数量比其余 3 种全局扰动的攻击方法少，也仅有本文攻击方法具有能让特定的分类器 100% 识别的可特定识别性。

6 结束语

本文提出了一种新的对抗攻击方法，该方法的特点是将密码学与对抗攻击相结合，使拥有密钥的分类器能够输出由本文方法生成的对抗样本的正确分类。这种识别方式可以保证特定分类对此类对抗样本具有高的识别能力。希望这一方法能够在未来的深度学习应用中提供一种新的安全性研究视角。

本文提出的对抗攻击方法的实例化采用了 RSA 方案。在保证密钥能同时运用在签名和加密方案的前提下，RSA 方案可以替换为其他密码学方案。更好的密码学方案可以让本文的对抗攻击安全性更高，且能够构造出长度更短的签名与密文信息，留出更多的空间用于计算扰动像素。这可以给扰动像素留出更多计算空间，让最终生成的扰动向量更加隐蔽和不明显。

在扰动噪声的位置上，本文方法选用的嵌入位置较为简单，未来可以从计算机视觉的角度在图像全局像素中寻找更优的嵌入位置，来实现更好的攻击效果。而限于密码学方案中对于密文信息和签名信息的完整性要求，如何为本文方法引入物理鲁棒性仍然是一个富有挑战性的问题。

参考文献：

- [1] LE-CUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
- [2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [3] GUO G, ZHANG N. A survey on deep learning based face recognition[J]. Computer Vision and Image Understanding, 2019, 189: 102805.
- [4] DENG L, HINTON G, KINGSBURY B. New types of deep neural network learning for speech recognition and related applications: an overview[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 8599-8603.
- [5] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep

- learning based natural language processing[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.
- [6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
- [7] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.
- [8] 何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全性问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049.
- HE Y Z, HU X B, HE J W, et al. Overview of the privacy and security issues of machine learning systems[J]. Computer Research and Development, 2019, 56(10): 2049.
- [9] ZHENG T, CHEN C, REN K. Distributionally adversarial attack[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 2253-2260.
- [10] XIAO C, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[J]. arXiv preprint arXiv:1801.02610, 2018.
- [11] QIU S, LIU Q, ZHOU S, et al. Review of artificial intelligence adversarial attack and defense technologies[J]. Applied Sciences, 2019, 9(5): 909.
- [12] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.
- [13] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial attacks and defences competition[M]//The NIPS'17 Competition: Building Intelligent Systems. 2018: 195-231.
- [14] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv preprint arXiv:1705.07204, 2017.
- [15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [16] LIANG B, LI H, SU M, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2018: 1-10.
- [17] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE Symposium on Security and Privacy (SP). 2016: 582-597.
- [18] JIA X, WEI X, CAO X, et al. Comdefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6084-6092.
- [19] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [20] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904.
- ZHANG S S, ZUO X, LIU J W, The problem of adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904.
- [21] 段广略, 马春光, 宋蕾, 等. 深度学习中对抗样本的构造及防御研究[J]. 网络与信息安全学报, 2020, 6(2): 1-11.
- DUAN G H, MA C G, SONG L, et al. Research on structure and defense of adversarial example in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(2): 1-11.
- [22] 刘西蒙, 谢乐辉, 王耀鹏, 等. 深度学习中的对抗攻击与防御[J]. 网络与信息安全学报, 2020, 6(5): 36-53.
- LIU X M, XIE L H, WANG Y P, et al. Adversarial attacks and defenses in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53.
- [23] 薛锐. 公钥加密理论[M]. 北京: 科学出版社, 2016.
- XUE R. Public key encryption theory[M]. Beijing: Science Press, 2016.
- [24] DIFFIE W, HELLMAN M. New directions in cryptography[J]. IEEE Transactions on Information Theory, 1976, 22(6): 644-654.
- [25] RIVEST R L, SHAMIR A, ADLEMAN L. A method for obtaining digital signatures and public-key cryptosystems[J]. Communications of the ACM, 1978, 21(2): 120-126.
- [26] DONG Y, FU Q A, YANG X, et al. Benchmarking adversarial robustness on image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 321-331.
- [27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [28] HYUN K, KIM Y, PARK K W, et al. Friend-safe evasion attack: an adversarial example that is correctly recognized by a friendly classifier[J]. Computers & Security, 2018, 78: 380-397.
- [29] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[作者简介]



张宇 (1995—), 男, 河南许昌人, 暨南大学博士生, 主要研究方向为深度学习安全、对抗样本。



李海良 (1981—), 男, 河南商丘人, 暨南大学副教授, 主要研究方向为图像识别与网络安全。