视觉对抗样本生成技术概述

王 伟[1], 董 晶[1], 何子文[1,2], 孙哲南[1]

[1]中国科学院自动化研究所智能感知与计算研究中心 北京 中国 100190
[2]中国科学院大学 北京 中国 100049

摘要 深度学习的发明, 使得人工智能技术迎来了新的机遇, 再次进入了蓬勃发展期。其涉及到的隐私、安全、伦理等问题也日益受到了人们的广泛关注。以对抗样本生成为代表的新技术, 直接将人工智能、特别是深度学习模型的脆弱性展示到了人们面前, 使得人工智能技术在应用落地时, 必须要重视此类问题。本文通过对抗样本生成技术的回顾, 从信号层、内容层以及语义层三个层面, 白盒攻击与黑盒攻击两个角度, 简要介绍了对抗样本生成技术, 目的是希望读者能够更好地发现对抗样本的本质, 对机器学习模型的健壮性、安全性和可解释性研究有所启发。

关键词 人工智能安全; 对抗样本; 白盒攻击; 黑盒攻击; 失真度量; 对抗防御
中图法分类号 TP301 DOI 号 10.19363/J.cnki.cn10-1380/tn.2020.02.04

A Brief Introduction to Visual Adversarial Samples

WANG Wei[1], DONG Jing[1], HE Ziwen[1,2] SUN Zhenan[1]

[1]Center for Research on Intelligent Perception and Computing, Institute of Automation,Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract With the invention of deep learning, artificial intelligence (AI) has ushered in new opportunities and is booming again. However, its privacy, security, ethics and other issues involved are also increasingly concerned by people. The adversarial samples, the vulnerability of artificial intelligence, especially deep learning models, are directly in front of us in recent years, which makes it necessary to pay attention to such problems during the practical application of AI technology. In this paper, a brief review of adversarial sample generation under white-box and black-box attack protocols is given. We summarize related techniques into three levels: signal level, content level and semantic level. We hope this paper can help readers better find the nature of the adversarial sample, which may improve the robustness, security and interpretability of the learned model.

Key words AI security; adversarial sample; white-box attack; black-box attack; distortion; adversarial defense

1 引言

随着深度学习技术[1]的快速发展, 其在公共安全、金融、医疗、娱乐等多个应用领域都取得了巨大成功, 人工智能技术再次迎来蓬勃发展期。以人脸识别[2]为代表的人工智能技术的快速落地, 尤其是在国内井喷式地发展, 给人们生产生活带来极大便利性的同时, 也引发了很多安全性问题[3]。

相较于传统安全问题, 如人脸识别系统中的照片攻击、假体攻击等, 一类新型的、具有攻击性的对抗样本生成技术近年来受到广泛关注, 特别是在视觉数据理解与分析领域。出于视觉数据特有的冗余特性, 一系列相关技术被公开发表, 并日益完善。鉴于此, 人们比以往任何时候都更加关注人工智能系统的安全问题, 对抗防御等应对技术也相继被提出。那么, 到底是何原因使得人工智能视觉系统易被攻击? 是深度神经网络模型的大规模应用带来的问题, 还是深度学习技术之前就已存在? 是个别系统特有, 还是绝大多数系统的共性问题? 若是后者, 目前是否有普适性的防御技术? 希望通过本文的介绍, 能够尽可能地回答上述问题。

本文首先针对人工智能机器学习中的安全问题展开宏观介绍及数学描述; 接着对具有代表性的对抗样本生成技术分类归纳介绍; 最后, 针对生物识别系统中人脸识别模型的安全性给出具体分析, 以便读者更为直观地了解对抗样本引发的安全问题。

通讯作者: 董晶, 博士, 副研究员, Email: jdong@nlpr.ia.ac.cn。
本课题得到国家自然科学基金 61972395、U1736119、61772529 资助。
收稿日期: 2020-01-03; 修改日期: 2020-02-20; 定稿日期: 2020-02-23

相较于本文, 文献[4]对抗样本经典方法进行了较为全面的介绍, 细致分析了对抗样本的产生原因、特征以及评价指标, 但其对近年来出现的新方法介绍略显不足。文献[5]极为详细的阐述了近来对抗样本的生成与防御方法, 其更多地是提供方法概览, 分析归纳对比的部分不足。本文主要贡献在于: 以视觉对抗样本生成技术为代表, 根据感知度量指标的不同定义, 挂一漏万地将对抗样本生成模型从信号层、内容层和语义层进行概括性的分类、总结和归纳, 并就对抗样本产生的原因、机理进行了深入分析。

2 机器学习中的安全问题

机器学习中的安全问题由来已久[6], 并不是深度学习时代所特有的。但是对抗样本的概念确是在深度学习广泛应用后被提出[7]。深度学习(特征学习)之前, 机器学习主要关注点是如何在已有的特征分布上学习到好的模型。对于模型的攻击往往是通过注入特定的训练样本(特征), 从而改变模型的决策界面, 实现攻击目的[6-7]。传统人工设计特征是在专家经验及人类对客观事物理解的基础上设计的, 具有较好的可解释性, 语义性较强, 对噪声和干扰鲁棒; 即输入样本的轻微扰动不会引起特征的变化, 从而不会改变模型的预测结果。而深度神经网络模型直接对输入和输出建模, 少了特征设计环节, 采用特征学习方式完成任务。其学习到的特征, 唯目标导向, 虽在性能上超越了传统方法, 但由于其黑盒特性, 学到的特征往往不具有可解释性, 从而使得对抗样本的生成较为容易[8-9]。这也就是深度学习时代对抗样本问题受到日益关注的原因。本文将重点对深度神经网络的对抗样本生成机理和具体算法展开介绍。

2.1 问题描述

针对某一任务的机器学习模型, 对抗样本是对输入样本轻微扰动后生成的, 其目的是改变模型的预测结果。对抗样本的求解过程可描述为:

$$\tilde{X} = \underset{\substack{D(X,\tilde{X})\le\epsilon;\\ X,\tilde{X}\in x}}{\arg\min} \left\{ \max P\left(f\left(\tilde{X}\right)|X\right) = \tilde{y} \right\} \tag{1}$$

其中, $f(X)\in\mathcal{Y}$ 为深度网络模型,它是在样本集合 $\{(X_i,y_i)\}_{i=\{1\cdots N\}}$上训练得到的, $\mathcal{X}$是样本空间, $\mathcal{Y}$是标签空间。$\tilde{X}$是在X基础上轻微扰动生成的, y为X的真实标签, $\tilde{y}$为$\tilde{X}$的预测标签。$D\left(X,\tilde{X}\right)$是对$\tilde{X}$和$X$的感知差异度量, 当$\tilde{y}\neq y$时, 我们称$\tilde{X}$为$X$的对抗样本[10]。

由公式(1)可知, $D\left(X,\tilde{X}\right)$对对抗样本的生成十分关键, 其上限$\epsilon$决定了对抗样本的攻击成功率和生成质量。目前大多数算法以人类视觉感知冗余度作为约束上限的, 如l_p范数[11]。由于其数学定义简单清晰, 常被用作度量不可感知性, 决定着对抗样本的攻击能力, 但其并不太符合人类视觉感知模型。因此, Jang 等人[12]提出了三个衡量标准来评估对抗样本不可感知性。通过对傅立叶系数的破坏程度、对边缘的影响或对图像梯度的影响来评估对抗扰动的质量。由于人类的视觉系统对结构变化很敏感, 结构相似性指数[13](SSIM)可作为内容相似性度量指标指导生成对抗样本[11]。在此基础上, 有学者提出使用感知对抗相似性分数[14](PASS)来量化对抗样本的不可感知性。进一步, 可借用 GAN 网络中的判别网络来评价生成对抗样本的质量[15]。从视觉冗余度到视觉合理性的改进, 增加了攻击强度空间。但随着对抗样本检测算法的提出[16], 感知上限ϵ势必会被大大压低, 即对抗样本的生成难度将会增加。

从应用角度, 防御方可通过设计对抗样本检测子[16], 压缩对抗样本的生成空间。然而, 对于防御方而言, 更重要的是如何提高模型$f(X)$鲁棒性。有文献表明, 可采用对抗训练[7]、模型正则约束[17]、主动防御[18]等策略来提高模型的对抗防御能力, 但其会降低模型的预测准确性。鉴于篇幅有限, 本文将不对防御工作展开详细介绍, 感兴趣的读者可参考文献[19]。

2.2 对抗样本生成

公式(1)给出了对抗样本的数学定义。那么如何计算求解最优的对抗样本? 最简单直观的算法就是采用穷举或搜索方法。然而, 对于视觉内容, 其搜索空间巨大, 不具备实际可操作性。我们知道, BP 算法的提出, 梯度下降法在神经网络模型训练过程中起到了至关重要的作用。对于模型f, 被正确预测样本X的后验概率$P(f(X)=y|X)$应该是最大的, 或者损失$L(f(X),y)$应该是最小的。如果一个样本$\tilde{X}$能够使得标签y的后验概率降低或损失增加, 直至小于其他标签的后验概率, 或大于其他标签的损失, 那么就可以说找到了对抗样本。对于靶向攻击, $\tilde{y}$作为对抗样本的预测标签, 其后验概率应是所有标签中最大的, 或者损失是所有标签中最小的。由此可见, 我们可以通过最大化模型损失函数(非靶向攻击)或最小化目标标签损失函数(靶向攻击)来生成对抗样本。

若令

$$L(f(X),\neg y) = -L(f(X),y) \tag{2}$$

且$\tilde{y}=\neg y$表示$\tilde{y}\neq y$, 则靶向攻击和非靶向攻击可统一在最小化损失框架下:

$$\min_{\substack{D\left(X,\tilde{X}\right)\le\epsilon;\\ X,\tilde{X}\in x}} L\left(f\left(\tilde{X}\right),\tilde{y}\right) \tag{3}$$

本文将结合感知差异度量指标$D\left(X,\tilde{X}\right)$, 从信号

层、内容层及语义层, 将对抗样本生成模型分为三类并分别介绍。

2.2.1 自适应加性噪声模型

若将对抗样本合理建模为原始信号上的加噪模型, 将其伪装为现实中常见的被噪声污染的正常样本, 即

$$\tilde{X} = X + \delta(X) \tag{4}$$

则对抗样本的求解过程将会变得较为容易。对抗样本不可感知性约束可简化为$D(X,\tilde{X}) = \left\|\tilde{X} - X\right\|_p = \|\delta(X)\|_p \leqslant \epsilon$, 即对加入噪声的$L_p$范数进行上限约束。与随机噪声不同的是, 加入的对抗噪声是自适应的, 与输入X有关。

当模型为线性函数$f(x) = w^{\mathrm{T}}x$, $\|\delta(x)\|_\infty \leqslant \epsilon$时, $\delta(x) = \epsilon sign(w)$是在约束条件下找到的最有可能的对抗样本[20]($f(x)$与$f(x+\delta(x))$差异最大)。然而, 对于高维样本数据, 线性模型往往性能较差, 现有主流方法主要采用高度非线性的深度神经网络模型$f(X)$, 实现从端到端的学习任务。我们知道, $L(f(X),y)$可在X_0附近一阶泰勒近似为: $L(f(X),y) \cong L(f(X_0),y) + \nabla_{\mathrm{X}} L(f(X_0),y)^{\mathrm{T}}(X - X_0)$。通过最小化$L(f(X),\tilde{y})$的泰勒近似, 可得$X_0$生成对抗样本$\tilde{X}_0$(在$X_0$的$l_\infty$邻域)的最优扰动为$-\epsilon sign\left(\nabla_{\mathrm{X}} L(f(X_0),\tilde{y})\right)$。然而, 由于泰勒近似的不准确性, 尤其是在高度非线性区域, 生成攻击成功的对抗样本所需的ϵ通常较大, 容易被发现。因此, 通过分段线性、小步迭代是自然而然的想法[21], 第 3 节将对此展开详细介绍。

2.2.2 几何扰动对抗模型

对不可感知性的度量, 除了 PSNR(可泛化到l_p范数)这类反映信号幅值变化的常见指标之外, 还可以用几何上的失真度量来限制视觉内容的改变。对于图像这类冗余度较高的视觉数据, 小范围的平移、旋转、缩放、变形等, 都不影响人类对视觉内容的感知。由此, 可定义基于几何扰动的对抗样本为[22]:

$$D(X,\tilde{X}) = \left\|\tilde{X}(r') - X(r)\right\|_p \leqslant \epsilon \tag{5}$$
$$r' \in \mathcal{B}_W(r,\delta) = \{r' | d_W(r,r') \leqslant \delta\}$$

其中$d_W(r,r')$为 Wasserstein 距离, 表示把一个分布变换为另一个分布最小代价。这里可将X归一化后看作是在位置r上的分布。通过在r为球心, δ为半径的 Wasserstein 球$\mathcal{B}_W(r,\delta)$上寻找最佳位移, 生成对抗样本$\tilde{X}$。此时, 对抗样本的求解过程可转化为关于位移场的优化问题。

2.2.3 语义保持对抗模型

前述的两种对抗样本生成模型是通过不同方法在视觉感知层面确保对抗样本与原始样本差异很小。如果突破该层面, 在更高的层面, 语义层上生成对抗样本[15], 将不再受限于信号层和内容层对修改容量的约束, 拥有更多的操作空间, 容易生成攻击成功率更高的对抗样本。该问题的难点是如何形式化判定不同样本属于同一个语义概念。目前而言, 该问题本身也一直是统计学习领域较难跨越的鸿沟。正是由于“感知-语义”鸿沟的存在, 为对抗样本的生成提供了便利, 使得相近内容得到不同语义理解成为可能。此时, 对生成样本的约束可表示为

$$D(X,\tilde{X}) = \left\|s(\tilde{X}) - s(X)\right\|_p \tag{6}$$

其中$\tilde{X} \in \mathcal{X}$要求在样本$X$的流形分布上。

由公式(6)可见, 相较于加性噪声模型对于扰动量上界的严格要求, 语义保持对抗模型对样本的修改更加自由。

3 对抗样本生成技术介绍

通过不同的距离度量, 我们可以在信号层、内容层和语义层, 多个层面实现对抗样本的计算, 发现智能系统的漏洞。信号层对抗样本生成是指仅从数据(信号)层面, 通过少量修改数据生成对抗样本而不考虑其对数据内容及含义的破坏。该类算法主要通过加性噪声模型实现。内容层对抗样本生成是指通过对数据内容的轻微修改生成对抗样本。内容上的修改可能会使数据数值发生较大变化, 可借助前述几何扰动对抗模型实现。语义层对抗样本生成较前两个层面的生成算法更为激进, 只要生成对抗样本不改变语义即可, 数值、内容均可发生较大变化。前述的语义保持对抗模型主要解决该问题。

当智能系统模型的参数可被获取时, 对抗样本的生成是较为容易的, 称为白盒攻击。然而, 大多数情况下, 模型参数是无法得到的, 只能获得预测标签, 称为黑盒攻击。此时对抗样本的生成是相对比较困难的。针对黑盒攻击, 除了利用有限次查询计算生成对抗样本之外, 还可利用对抗样本的迁移特性[23], 对替代模型实施白盒攻击, 从而实现对部署模型的黑盒攻击。本节中我们将对代表性的白盒和黑盒攻击算法展开介绍。

3.1 白盒攻击

白盒攻击是指攻击者在拥有目标模型的结构和参数等全部知识的情况下对模型实施攻击。白盒攻击一般依赖于模型的梯度信息, 利用模型损失函数反向传播的梯度计算扰动量, 将其添加到原始样本上形成对抗样本。

Szegedy 等[7]是最早发现神经网络存在对抗攻击的研究者之一, 他们提出通过L-BFGS优化方法去解

决一个优化问题来找到最小的可能的攻击扰动, 缺点在于复杂的优化过程使得运算速度很慢。Ian Goodfellow 等[20]提出了快速梯度符号攻击方法(Fast Gradient Sign Method, FGSM), 极大地加速了对抗样本的生成。其主要思想是寻找分类模型的梯度变化最大的方向, 根据此方向添加图像扰动, 导致模型进行错误的分类。FGSM 以增加对图像分类器损失的方式来对图像添加扰动, 其构造对抗样本的优势是效率高, 算法不需要利用梯度的全部信息, 只需要判定梯度符号即可。FGSM 作为经典的攻击方式, 衍生出了许多以其为基础的对抗攻击方法。Kurakin 等[21]提出了一种迭代化的 FGSM, 称为基础迭代方法(Basic Iterative Method, BIM)。FGSM 只沿着梯度增加的方向添加一步扰动, 而 BIM 则通过迭代的方式, 沿着梯度增加的方向进行多步小扰动, 并且在每一小步后, 重新计算梯度方向, 相比 FGSM 能构造出更加精准的扰动, 但代价是增大了计算量。这种方法基于 FGSM, 因此也常称为 I-FGSM。Madry 等[24]提出了投影梯度下降(Projected Gradient Descent, PGD)生成对抗样本, 用于对抗训练的方法, 较好地提高了模型防御能力。PGD 是在 BIM 的基础上, 对原始样本在其邻域范围内随机扰动作为算法初始输入, 经多次迭代后生成对抗样本, 其性能得到显著改善, 具有较好的迁移性和抗破坏能力。

FGSM 及其衍生算法相较于 L-BFGS[7]极大地提升了攻击成功率以及运算速度, 但并未对扰动量直接优化求解。Seyed-Mohsen 等[25]提出了一种新的迭代攻击方式 Deepfool, 具有 FGSM 类似的欺骗率的同时能够计算出更小的扰动。该方法首先初始化原始图像, 并且假定该图像类别由分类器的决策边界限制在一个区域, 这个区域决定图像的类标签。在每次迭代中, 该算法计算幅值较小的向量来扰动图像, 通过线性化图像所在区域的边界, 逐步将图像移向决策边界, 直到图像最终被移动到决策边界另一侧, 使得分类器分类错误。N.Carlini 和 D.Wagner[26]进一步改善了 L-BFGS 的攻击性能, 提出了 C&W 攻击。该方法巧妙地设计了损失函数, 使其在对抗样本中有较小的值, 但在干净样本中有较大的值, 进而通过最小化该损失函数即可搜寻到对抗样本。与 L-BFGS 方法另一个不同之处在于, 其使用 Adam[27]算法求解优化问题, 并在每步迭代之后将结果投影到边界约束, 继而解决边界约束的问题。相较于前述算法, 其速度更快, 效果更好。C&W 攻击中的优化项需要进行上千次的线搜索迭代, 造成运算时间较长, Jerome 等人进一步对其完善, 提出了 Decoupling Direction and Norm Attack(DDNAttack)[28]算法, 通过优化交叉熵目标函数实现。不同于传统梯度迭代算法在每次迭代时惩罚正则项, DDNAttack 在每次迭代中先根据梯度方向添加扰动, 然后将得到扰动映射到以原始样本为中心的l_2球上, 球半径由可控, 其值是否修改由当次迭代样本是否为攻击成功决定。通过解偶扰动方向与幅值, DDNAttack 大量地减少了迭代次数, 但仍能得到效果与 C&W 相当的对抗样本。

为了防御基于梯度的白盒攻击方法, 一些防御使用混淆梯度[29]的策略来干扰攻击过程, 然而梯度混淆防御策略很容易被攻破。攻击者可以采用反向传播微分近似算法(Backward Pass Differentiable Approximation, BPDA)[29], 针对防御者添加的不可微预处理模块, 获取近似梯度, 进行实现有效攻击; 当防御者采用随机化梯度方式时, 攻击者亦可利用求随机变换期望(Expectation over Transformation, EOT)的替代方法[30]实施攻击。

除采用常规的l_∞范数或l_2范数的对抗攻击模式外, 许多攻击方法也特意地研究了其他模式的对抗扰动。Papernot 等提出了基于雅克比计算显著图的攻击方法(Jacobian-based Saliency Map Attack, JSMA)[31]。在图像识别问题中, 不同像素点对识别结果的重要性显然是不同的, JSMA 通过计算图像显著性分数, 依照重要程度, 逐个对单像素进行修改得到对抗样本。因此, 可通过仅改变部分图像像素实现有效攻击。Xiao 等[32]提出了一种基于图像内容几何变换而非直接修改像素值的方法来生成对抗样本, 该几何变换是基于对原始图像中的像素进行位移, 将其限制为全局平滑。通过对误分类概率优化求解对抗样本。类似地, Alaifari 等[33]提出一种通过图像形变(内容保持)来构建对抗样本的迭代算法 ADef, 类似于迭代梯度算法, 该算法根据梯度下降方向通过迭代地对图像进行变形操作从而将原始样本推向决策边缘以获得对抗样本。

Song 等[15]提出一种基于生成式对抗网络(Generative Adversarial Networks, GAN)[34]的方法, 生成语义保持的对抗样本, 这类对抗样本也被称作非受限样本。

3.2 黑盒攻击

黑盒攻击是指攻击者在无法获取到目标模型结构或参数等内部信息的情况下对模型进行攻击。黑盒攻击大致可分为两类方法, 基于查询的方法和基于模型迁移的方法。

3.2.1 基于查询的方法

在实际应用中许多模型不公开其内部结构而只

提供调用接口, 例如人脸识别 API 等。基于查询的攻击方法则通过不断地访问目标模型, 近似估计模型梯度信息, 以达到修改输入, 生成对抗样本的目的。

Chen 等[35]提出了零阶优化(Zeroth Order Optimization, ZOO)来估计目标模型的梯度以产生对抗样本。其假设目标模型能够被查询以获得所有类的概率得分, 然后使用差分数值近似估计目标函数关于输入的梯度, 进而使用基于梯度的方法进行攻击。类似地, Jonathan 等[36]提出了一种使用同时扰动随机逼近算法(Simultaneous Perturbation Stochastic Approximation, SPSA)进行梯度估计实施攻击的方法, 通过特征降维以及随机抽样比 ZOO 取得了更高的效率。Tu 等[37]提出了一种高效的黑盒攻击框架 AutoZOOM (Autoencoder-based Zeroth Order Optimization Method), 使用一种自适应随机梯度估计方法稳定查询次数和扰动量, 同时利用无标签数据离线训练自动编码器, 从而加快对抗样本的生成。Brendel 等[38]提出了一种基于边界探索的黑盒攻击方式, 称为边界攻击(Boundary Attack)。这种方法完全依赖于模型的最终决策输出(例如 Top1 类别标签), 其所需知识更少。该方法首先寻找到一个不限制扰动大小的对抗样本, 然后依据一定策略将该对抗样本沿着原样本的方向移动, 直到该对抗样本离原样本最近, 其结果依然保持较强的对抗性。

在实际应用中, 为了不被察觉, 对目标模型的访问查询是受限的, 包括访问时间限制与访问次数限制。Ilyas 等[39]针对该问题, 提出了一种自然进化策略(Natural Evolution Strategies, NES)的变体, 并结合投影梯度下降(PGD), 构建对抗样本。Bhagoji 等[40]采用随机置k个元素为1, 其余为 0 的方式, 产生$\lceil d/k \rceil$组相互正交的随机向量, 通过方向导数与梯度关系, 结合使用主成分分析(Principal Component Analysis, PCA), 近似得到最终的梯度估计, 从而将查询次数降低k倍。Chen 等[41]进一步对边界攻击[38]进行改进, 提出了 HopSkipJumpAttack。其基于在决策边界处使用二值信息进行梯度方向估计, 显著减少了对目标模型的查询次数。

此外, 还有学者对黑盒攻击扰动的稀疏性进行了探索。Su 等[42]提出了单像素攻击(One Pixel Attack)方法, 利用差分进化算法从候选像素点中逐步筛选出稀疏像素点, 最终修改选择的像素点可以有效攻击分类模型。Narodytska 等[43]提出了局部搜索攻击(Local Search Attack), 通过贪心局部搜索计算选取待修改像素点。Apostolos 等[44]提出了 SparseFool, 一种受几何启发的稀疏攻击, 利用边界的低平均曲率有效地计算对抗性扰动。该方法能够快速地计算稀疏扰动, 且可以有效拓展到高维数据上。

3.2.2 基于模型迁移的方法

尽管基于查询的方法十分有效, 但需对目标模型进行大量访问, 易被察觉。基于模型迁移的方法很好的弥补了该缺点。它利用对抗样本在同一任务的不同模型之间具有迁移性的特点[23], 通过攻击同一任务的替代模型生成对抗样本。因此, 对抗样本的迁移性显得尤为重要。基于模型迁移的方法在不访问目标模型的情况下便可成功实现黑盒攻击。

Papernot 等[23]最早发现对抗样本的迁移性: 攻击一个模型所生成的对抗样本往往能够同时欺骗同一任务其他模型。之后产生了大量提升对抗样本迁移性的工作。为了解决大多数对抗攻击迁移性较差, 黑盒攻击成功率较低的问题, Dong 等提出了一种基于动量的迭代算法 MI-FGSM[45]。通过将动量项集成到攻击的迭代过程中, 来增强对抗样本迁移性。动量项在迭代过程中起到稳定更新方向、避免局部极值的作用, 从而产生迁移性更强的对抗样本。Curls & Whey[46]在 MI-FGSM 基础上进行改进, 通过结合梯度上升和梯度下降的方向, 增加迭代轨迹的多样性, 生成迁移性更强的对抗样本, 同时还提出了利用扰动的鲁棒性来去除冗余噪声, 实现了在相同扰动量的情况下迁移性更强的目的。Xie 等[47]利用对输入的多种扰动来增强所生成对抗样本的迁移性。其通过在 BIM 算法基础上每次迭代时对输入做随机变换来实现输入的多样化, 该操作使得生成的对抗噪声与内容更相关, 更难被破坏。实验发现随机尺度变换和随机零填充结合的方法效果最优。类似地, Dong 等[48]提出了一种平移不变的攻击方法来生成迁移性更强的对抗样本。通过优化一组平移图像的扰动生成对抗样本, 提升其迁移性。为了提高攻击的效率, 通过将原图像的梯度与一个预定义的核函数卷积实现平移操作效果, 提升了方法效率, 同样可应用在其他基于梯度的攻击方法上。Liu 等[49]提出将一系列不同架构的模型集成作为替代模型。针对该集成替代模型生成的对抗样本极大地增强了攻击迁移性, 在黑盒攻击中表现明显好于单一替代模型。然而, 针对集成模型的黑盒攻击算法, 其性能往往受到模型数量限制。Li 等[50]提出构建多样性集成模型的算法, 通过在已有模型基础上进行特征级扰动, 构建一组多样化模型用于集成, 从而提高对抗样本迁移性。实验结果表明, 增加集成模型数目对提高对抗样本迁移性是至关重要。因此, Che 等[51]提出了一种新的黑盒攻击, 称为连环小批量集成攻击, 核心在于将预先

训练好的大量源模型分成几个小批量, 并将前一个批次梯度长期记忆递归累积至下一个批次, 从而提升可迁移性。上述集成方法, 主要通过直接平均各模型输出来构造一个简单的集成, 忽略了模型间的相关性。Pang 等[52]通过研究各参与集成模型间的相关性, 利用正则化项约束, 自适应地促进集成模型的多样性, 提高对抗样本的迁移性和鲁棒性。

3.3 物理空间对抗样本

目前, 大多数对抗样本生成研究主要集中在数字空间。那么, 这些对抗样本在三维真实物理环境中是否仍然有效? 已有相关研究表明, 物理可实现的对抗样本在真实世界中对机器学习模型依然具有一定的威胁, 但可想而知, 其生成困难, 且易被察觉。目前主要研究思路是将物理空间到数字空间的感知过程用函数近似, 例如将视觉成像过程由射影变换与颜色空间变换模拟。目的是使对抗样本与真实样本在成像后的数字空间中足够相似, 即对抗样本生成过程中应当考虑成像过程对对抗噪声的破坏。因此, 物理空间的对抗样本需要更大失真才有可能成功攻击智能系统。本质而言, 物理空间的攻击基本上都是采用前述在数字空间中使用的策略和方法, 其难点在于对物理空间到数字空间变换过程的可微模型建立。

Kurakin 等[21]研究了数字空间利用 FGSM 算法生成的对抗样本经过打印后在物理空间的对抗性能。实验发现, 所生成对抗样本较为脆弱, 较大地受限于光照、距离、角度等物理成像条件。Eykholt 等[53]提出了一种鲁棒物理扰动(Robust Physical Perturbation, RP_2)计算方法, 其能够在角度、距离和光线变化时保持攻击的成功率。作者认为实际攻击应该发生在目标上且不易引起人关注的区域, 掩膜(Mask)的引入很好的解决了该问题。通过实际采集和数据合成的方式, 实现了对物理成像过程的统计建模。最终结合颜色打印损失模型, 实现对物理空间中停止标志的攻击。作者利用该攻击算法生成并打印制作了路标图像, 成功欺骗了实际部署应用的多数路标分类模型。文中选用路标做实验, 其实是将问题难度进行了简化。路标为平面物体, 因此三维成像过程只有仿射变换, 省去了射影变换中的二义问题。因此, 在其他非平面目标上的实验性能如何, 有待实验论证。针对上述问题, Athalye 等[30]考虑更复杂的空间变换关系。作者首次提出了 EOT(Expectation Over Transformatio)算法, 使得对抗样本在成像过程中多种变换下都有效。在该框架下结合 3D 打印技术生成了 3D 对抗样本, 成功欺骗了目标分类识别系统, 从不同角度观察时均使目标模型分类错误。多数针对智能视觉系统的攻击都尽力使对抗目标与原始目标成像后在数字空间保持一致。不同于此, 一类对抗补丁(Adversarial Patch)方法, 其目的只是攻击识别系统, 并不考虑补丁的不可察觉性。Brown 等[54]提出了一种对抗补丁生成算法来欺骗分类器。在优化目标中仅考虑对抗补丁的放置位置及可能经过的变换, 因此生成的对抗补丁具有较强的攻击性, 但不具有不可察觉性。基于相同原理, Thys 等[55]通过在身体上悬挂对抗补丁, 成功欺骗了 YOLOv2[56]目标检测系统。类似地, Komkov 等[57]通过空间变换关系制作了一种物理空间的对抗贴纸。通过将对抗贴纸贴在帽子上, 成功地欺骗了包括 ArcFace[2]在内的最先进的人脸识别系统, 其攻击过程仍是在 FGSM 框架下。

4 生物识别系统的对抗样本

为了能够对对抗样本有直观的了解, 本节将针对典型的、应用较为广泛的人脸识别系统, 采用上述提到的几类攻击算法的开源代码生成人脸识别系统的对抗样本。

针对目前主流的人脸识别模型: ArcFace[2], 我们进行了白盒攻击和黑盒攻击实验。实验中白盒攻击的目标模型为 ArcFace。黑盒攻击中基于查询访问的方法, 目标模型为 ArcFace; 对基于模型迁移的黑盒攻击方法, 我们利用 ArcFace 作为替代模型离线生成对抗样本攻击 CosFace[58]模型。

我们从 LFW 数据集[59]中随机选取两张人脸图像, 一张作为待攻击样本(人脸图像 A), 一张作为攻击目标(人脸图像 B), 进行靶向攻击。我们选取了前面介绍的 7 种攻击方法: FGSM[20]、BIM[21]、MI-FGSM[45]、SPSA[36]、HopSkipJumpAttack[41]、DI-FGSM[47]、TI-FGSM[48]。对 ArcFace 模型, 当模型提取的人脸特征间欧式距离≤1.521时, 两张人脸图像被判定为同一个人。对 CosFace 模型, 当两个人脸特征的余弦距离≤0.2245时被判定为同一个人。

表 1 显示的是攻击模型成功时所需的最小扰动量。由表 1 可知, 当对模型攻击成功时, 白盒攻击相对比较容易, 且所需的扰动也比较小。但黑盒攻击除了在方法上需要特殊考量外, 其攻击强度相较于白盒攻击也要大出许多。FGSM 对抗样本生成方法, 甚至在黑盒攻击下完全失败。对黑盒攻击, 基于查询访问的方法所需的扰动量较小, 但由于其查询次数较多, 容易被察觉。而基于模型迁移的方法虽然可以离线完成对抗样本的生成, 实现一次性攻击, 但其攻击强度要大很多。

表 1 典型方法靶向攻击成功时的最小扰动量比较

Table 1 TargetedadversarialSampleswith minimal distortion

攻击类型		方法	最小扰动(l_∞)
白盒	基于梯度	*FGSM*[20]	0.024
		BIM[21]	0.022
		MI-FGSM[45]	0.024
	基于访问	*SPSA*[36]	0.020
		HopSkipJumpAttack[41]	0.031
黑盒	基于迁移	*FGSM*[20]	∞
		BIM[21]	0.35
		MI-FGSM[45]	0.22
		DI-FGSM[47]	0.18
		TI-FGSM[48]	0.16

注 1: 表中扰动量是将图像像素值归一化到[0,1]后计算得到。

注 2: ∞表示攻击失败。

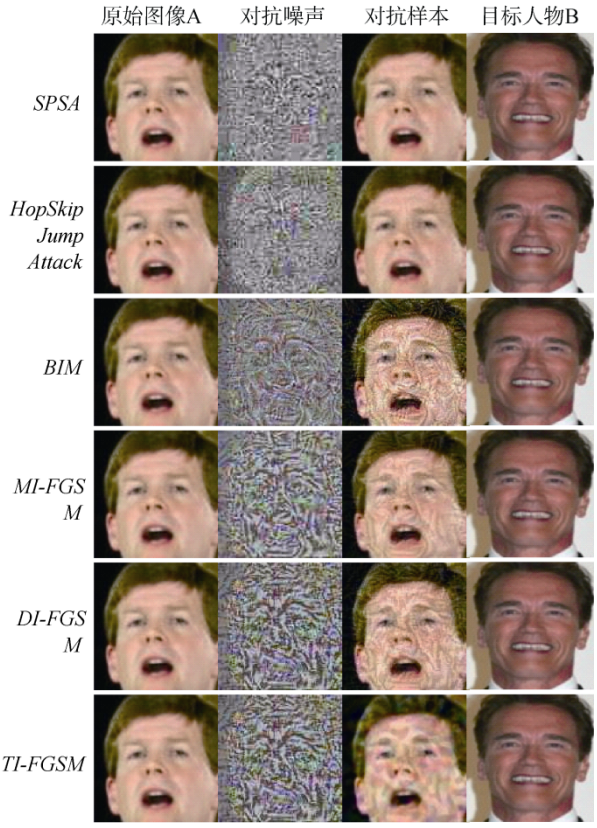


图 1 不同黑盒攻击算法生成的靶向对抗人脸图像

Figure 1 Targeted adversarialfaceimagesunder different black-box attack methods

由图 1 可看出, 黑盒攻击的成功率是以牺牲对抗样本的不可感知性为代价的。而对于采用查询访问部署模型的方法, 其能够较为准确地近似梯度信息, 类似白盒策略可获得较低的样本失真, 但此类方法查询次数较多, 攻击行为易被发现。

5 总结与展望

随着人工智能技术的大规模普及, 其涉及到的隐私、安全、伦理等问题受到人们的广泛关注。以对抗样本生成为代表的新技术, 直接将人工智能、特别是深度学习模型的脆弱性展现到了人们面前, 使得此类技术在应用落地时, 不得不重视其安全问题。

实际上, 对抗样本生成技术的研究目的更重要地是发现模型漏洞, 对模型的外延不断试探, 通过对抗训练等防御方法, 提高模型健壮性。另一方面也是对以卷积神经网络为代表的深度神经网络模型的可解释性进行尝试性研究, 加固模型, 减少数据的不可解释现象, 赋予实际意义。

本文从信号层、内容层以及语义层三个层面, 白盒攻击与黑盒攻击两个角度, 对对抗样本生成技术的展开了简要介绍, 目的是希望读者能够更好地发现对抗样本本质。然而, 仍有许多必须要面对的现实问题: 对抗样本的迁移能力到底有多大? 针对某一系统设计的对抗样本, 在同类任务不同模型上是否奏效[47]? 对抗样本健壮性如何[53]? 是否能够被轻易破坏[60-61]? 是否存在强安全模型, 使得对抗样本的生成非常困难[17]? 除了对抗训练方法外, 有没有更高效的模型加固方法[62]? 目前, 模型加固方案仍然具有较大局限性, 未能较好地解决安全问题。到底是什么原因使得深度神经网络模型深受对抗样本困扰[63]? 目前来说, 仍未有明确答案。

参考文献

[1] HintonGE, Osindero S,TehYW. A Fast Learning Algorithm for Deep Belief Nets[J].*Neural Computation*, 2006, 18(7): 1527-1554.

[2] Jiankang Deng, Jia Guo, Niannan Xue et al. ArcFace: Additive Angular Margin Loss for Deep Face Recognition[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 4690–4699.

[3] Yingzhe He, Guozhu Meng, Kai Chen et al. Towards Privacy and Security of Deep Learning Systems: A Survey. 2019: *arXiv preprint arXiv:1911.12562*, Nov..

[4] Zhang JL, Li C. Adversarial Examples: Opportunities and Challenges[J].*IEEE Transactions on Neural Networks and Learning Systems*, 2019: 1-16.

[5] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia et al. Adversarial Examples in Modern Machine Learning: A Review. 2019: *arXiv preprint arXiv:1911.05268*.

[6] Gregory L. Wittel , S. Felix Wu, On Attacking Statistical Spam Filters[C]. *CEAS: First Conference on Email and Anti-Spam*, 2004:34-41.

[7] Christian Szegedy *et al.*, Intriguing Properties of Neural Networks[C]. *2nd International Conference on Learning Representations, ICLR 2014-Conference Track Proceedings*, 2014: 23-41.

[8] Ali Shafahi *et al.*, Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks[C].*Advances in Neural Information Processing Systems*, 2018:56-61.

[9] Ji Feng, Qi-Zhi Cai , Zhi-Hua Zhou, Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder[C]. *Advances in Neural Information Processing Systems*, 2019: 11971-11981.

[10] Hyun Kwon, Yongchul Kim, Ki Woong Park et al. Multi-Targeted Adversarial Example in Evasion Attack on Deep Neural Network," *IEEE Access*, 2018:456-461.

[11] Sharif M, Baue L,Reite M K. On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June18-22, 2018. Salt Lake City, UT, USA. Piscataway, NJ: IEEE, 2018: 1686-1694.

[12] Jang U, Wu X, Jha S. Objective Metrics and Gradient Descent Algorithms for Adversarial Examples in Machine Learning[C]. *Proceedings of the 33rd Annual Computer Security Applications Conference on - ACSAC 2017*, December 4-8, 2017. Orlando,FL, USA. New York, USA: ACM Press, 2017: 262-277.

[13] Wang Z, Bovik A C, Sheikh H R,etal. Image Quality Assessment: From Error Visibility to Structural Similarity[J].*IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.

[14] Rozsa A, RuddEM,BoultT E. Adversarial Diversity and Hard Positive Generation[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 26-July 1, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 410-417.

[15] Yang Song, Nate Kushman, Rui Shu et al. Constructing Unrestricted Adversarial Examples with Generative Models[C]. *Advances in Neural Information Processing Systems*, 2018: 8312-8323.

[16] Liu J Y, Zhang W M, Zhang Y W, et al. Detection Based Defense Against Adversarial Examples from the Steganalysis Point of View[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 4825-4834.

[17] Hongyang Zhang, Yaodong Yu, Jiantao Jiao et al. Theoretically Principled Trade-off between Robustness and Accuracy[C]. *International Conference on Machine Learning*, 2019: 7472–7482.

[18] Hadi Salman et al. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers[C]. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019: 11289-11300.

[19] Wang XM, LiJ, KuangXH,etal. The Security of Machine Learning in an Adversarial Setting: A Survey[J].*Journal of Parallel and Distributed Computing*, 2019, 130: 12-23.

[20] Ian J. Goodfellow, Jonathon Shlens , Christian Szegedy, Explaining and Harnessing Adversarial Examples[C]. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015:456-461.

[21] Alexey Kurakin, Ian J. Goodfellow , Samy Bengio, Adversarial Examples in the Physical World[C]. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017:23-31.

[22] Eric Wong, Frank R. Schmidt , J. Zico Kolter, Wasserstein Adversarial Examples via Projected Sinkhorn Iterations[C], *the 36th International Conference on Machine Learning (ICML)*, 2019: 6808-6817.

[23] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, et al. Practical Black-Box Attacks against Machine Learning[C]. *ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, 2017: 506-519.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C] *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018z;34-41.

[25] Moosavi-Dezfooli SM,FawziA, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 2574-2582.

[26] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy (SP)*, May 22-26, 2017. San Jose, CA, USA. Piscataway, NJ: IEEE, 2017: 39-57.

[27] Diederik P. Kingma , Jimmy Lei Ba, Adam: A Method for Stochastic Optimization[C]. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015:23-41.

[28] Rony J, HafemannLG, OliveiraLS, etal. Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 4322-4330.

[29] Anish Athalye, Nicholas Carlini , David Wagner, Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples[C]. *35th International Conference on Machine Learning, ICML 2018*, 2018: 436-448.

[30] Anish Athalye, Logan Engstrom, Andrew Ilyas et al. Synthesizing Robust Adversarial Examples[C]. *35th International Conference on Machine Learning, ICML 2018*, 2018: 449–468.

[31] Papernot N, McDanielP, JhaS, etal. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, March 21-24,2016. Saarbrucken.Piscataway, NJ:IEEE, 2016: 372-387.

[32] Chaowei Xiao, Jun Yan Zhu, Bo Li, et al. Spatially Transformed Adversarial Examples[C]. *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, 2018: 23-31.

[33] Rima Alaifari, Tandri Gauksson , Giovanni S. Alberti, ADEF: An Iterative Algorithm to Construct Adversarial Deformations[C]. *7th International Conference on Learning Representations, ICLR 2019*, 2019: 45-51.

[34] Ian J. Goodfellow et al. CausalCategorizationwithBayes Nets[M]// Advances in Neural Information Processing Systems 14. The MIT Press, 2002.

[35] Pin Yu Chen, Huan Zhang, Yash Sharma, et al. ZOO: Zeroth Order Optimization Based Black-Box Atacks to Deep Neural Networks

without Training Substitute Models[C] *AISec 2017 - Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*,2017: 15–26.

[36] Jonathan Uesato, Brendan O'Donoghue, Aaron Van Den Oord et al. Adversarial Risk and the Dangers of Evaluating against Weak Attacks[C]. *35th International Conference on Machine Learning, ICML 2018*, 2018: 5032–5041.

[37] Tu C C, Ting P S, Chen P Y, et al. AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33: 742-749.

[38] Wieland Brendel, Jonas Rauber, Matthias Bethge, Decision-Based Adversarial Attacks: Reliable Attacks against Black-Box Machine Learning Models[C]. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018: 87-89.

[39] Andrew Eyas, Logan Engstrom, Anish Athalye et al. Black-Box Adversarial Attacks with Limited Queries and Information[C]. *35th International Conference on Machine Learning, ICML 2018*, 2018: 3392–3401.

[40] Bhagoji AN, He W, Li B, et al. Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms[M]. Computer Vision–ECCV2018.Cham:SpringerInternationalPublishing, 2018: 158-174.

[41] Jianbo Chen, Michael I. Jordan , Martin J. Wainwright, HopSkipJumpAttack: A Query-Efficient Decision-Based Attack, 2019: *ArXiv*.

[42] Su JW, Vargas DV, Sakurai K. One Pixel Attack for Fooling Deep Neural Networks[J].*IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828-841.

[43] Nina Narodytska, Shiva Prasad Kasiviswanathan, Simple Black-Box Adversarial Perturbations for Deep Networks, 2016: *ArXiv*.

[44] Modas A, Moosavi-Dezfooli SM, Frossard P. SparseFool: A few Pixels Make a Big Difference[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 9087-9096.

[45] Yinpeng Dong, Fangzhou Liao, Tianyu Pang *et al.* Boosting Adversarial Attacks with Momentum[C]. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018: 9185–9193.

[46] Shi Y C, Wang SY, Han Y H. Curls & Whey: Boosting Black-Box Adversarial Attacks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 6519-6527.

[47] Xie C H, Zhang Z S, Zhou Y Y, et al. Improving Transferability of Adversarial Examples with Input Diversity[C]. 2*019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 2730-2739.

[48] Dong YP, Pang T, Su H, et al. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 4312-4321.

[49] Yanpei Liu, Xinyun Chen, Chang Liu, et al. Delving into Transferable Adversarial Examples and Black-Box Attacks[C]. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019: 1-7.

[50] Yingwei Li, Song Bai, Yuyin Zhou, et al. Learning Transferable Adversarial Examples via Ghost Networks, 2018: *arXiv preprint arXiv:1812.03413*.

[51] Zhaohui Che, Ali Borji, Guangtao Zhai, et al. A New Ensemble Adversarial Attack Powered by Long-Term Gradient Memories. 2019: *arXiv preprint arXiv:1911.07682*.

[52] Tianyu Pang, Kun Xu, Chao Du, et al. Improving Adversarial Robustness via Promoting Ensemble Diversity[C]. *36th International Conference on Machine Learning, ICML 2019*, 2019: 8759–8771.

[53] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Visual Classification[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018. Salt Lake City, UT, USA. Piscataway, NJ: IEEE, 2018: 1625-1634.

[54] Tom B. Brown, Dandelion Mané, Aurko Roy, et al. Adversarial Patch. 2017: *arXiv preprint arXiv:1712.09665*.

[55] Simen Thys, Wiebe Van Ranst , Toon Goedemé, Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection[C]. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 34-51.

[56] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017. Honolulu, HI. Piscataway, NJ: IEEE, 2017: 6517-6525.

[57] Stepan Komkov , Aleksandr Petiushko, AdvHat: Real-World Adversarial Attack on ArcFace Face ID System. 2019: *arXiv preprint arXiv:1908.08705*.

[58] Wang H, Wang Y T, Zhou Z, et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018. Salt Lake City, UT. Piscataway, NJ: IEEE, 2018: 5265-5274.

[59] Lina J. Karam , Tong Zhu, Quality Labeled Faces in the Wild (QLFW): A Database for Studying Face Recognition in Real-World Environments[C]. *Human Vision and Electronic Imaging* ,2015: 345-351.

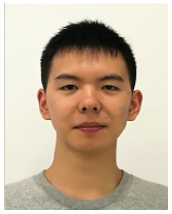
[60] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, *et al.*, Shield: Fast, Practical Defense and Vaccination for Deep Learning Using JPEG Compression[C]. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018: 196–204.

[61] Liao F Z, Liang M, DongY P, et al. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018. Salt Lake City, UT. Piscataway, NJ: IEEE, 2018: 1778-1787.

[62] Xie C H, WuY X, van der Maaten L, et al. Feature Denoising for Improving Adversarial Robustness[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 501-509.

[63] Adi Shamir, Itay Safran, Eyal Ronen, et al. A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance. 2019: *arXiv preprint arXiv:1901.10861*.

王伟 于 2012 年在中国科学院大学计算机应用技术专业获得博士学位。现任中国科学院自动化研究所副研究员。研究领域为人工智能安全、媒体内容取证与安全、隐写分析。研究兴趣包括: 对抗样本生成与防御、Deepfake 检测、篡改取证。Email: wwang@nlpr.ia.ac.cn



何子文 于 2018 年在上海交通大学自动化专业获得学士学位。现在中国科学院自动化所攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 对抗攻击与防御。Email: ziwen.he@cripac.ia.ac.cn

董晶 于 2010 年在中国科学院大学模式识别与智能系统专业获得博士学位。现任中国科学院自动化研究所副研究员。研究领域为人工智能安全、媒体内容取证与安全、隐写分析。研究兴趣包括: 隐写与隐写分析、Deepfake 检测、篡改取证。Email: jdong@ nlpr.ia.ac.cn

孙哲南 于 2006 年在中国科学院自动化研究所获模式识别与智能系统专业博士学位。现任中国科学院自动化研究所副总工程师、研究员和博士生导师、天津中科智能识别产业技术研究院院长。研究领域为生物特征识别与安全、模式识别、计算机视觉。研究兴趣包括: 虹膜识别、人脸识别、Deepfake 检测。Email: znsun@nlpr.ia.ac.cn