自动化学报
ACTA AUTOMATICA SINICA

车牌识别系统的黑盒对抗攻击

陈晋音 沈诗婧 苏蒙蒙 郑海斌 熊晖

Black-box Adversarial Attack on License Plate Recognition System

CHEN Jin-Yin, SHEN Shi-Jing, SU Meng-Meng, ZHENG Hai-Bin, XIONG Hui

在线阅读 View online: https://doi.org/10.16383/j.aas.c190488

您可能感兴趣的其他文章

基于贝叶斯序贯博弈模型的智能电网信息物理安全分析

Cyber-physical Security Analysis of Smart Grids With Bayesian Sequential Game Models

自动化学报. 2019, 45(1): 98-109 https://doi.org/10.16383/j.aas.2018.c180336

面向电力信息物理系统的虚假数据注入攻击研究综述

A Review on False Data Injection Attack Toward Cyber-physical Power System

自动化学报. 2019, 45(1): 72-83 https://doi.org/10.16383/j.aas.2018.c180369

网络攻击下信息物理融合电力系统的弹性事件触发控制

Resilient Event-triggered Control of Grid Cyber-physical Systems Against Cyber Attack

自动化学报. 2019, 45(1): 110-119 https://doi.org/10.16383/j.aas.c180388

假数据注入攻击下信息物理融合系统的稳定性研究

On the Stability of Cyber-physical Systems Under False Data Injection Attacks

自动化学报. 2019, 45(1): 196-205 https://doi.org/10.16383/j.aas.2018.c180331

基于条件深度卷积生成对抗网络的图像识别方法

Image Recognition With Conditional Deep Convolutional Generative Adversarial Networks

自动化学报. 2018, 44(5): 855-864 https://doi.org/10.16383/j.aas.2018.c170470

车牌识别系统的黑盒对抗攻击

陈晋音 [1] 沈诗婧 [1] 苏蒙蒙 [1] 郑海斌 [1] 熊 晖 [1]

摘 要 深度神经网络 (Deep neural network, DNN) 作为最常用的深度学习方法之一, 广泛应用于各个领域. 然而, DNN 容易受到对抗攻击的威胁, 因此通过对抗攻击来检测应用系统中 DNN 的漏洞至关重要. 针对车牌识别系统进行漏洞检测, 在完全未知模型内部结构信息的前提下展开黑盒攻击, 发现商用车牌识别系统存在安全漏洞. 提出基于精英策略的非支配排序遗传算法 (NSGA-II) 的车牌识别黑盒攻击方法, 仅获得输出类标及对应置信度, 即可产生对环境变化较为鲁棒的对抗样本, 而且该算法将扰动控制为纯黑色块, 可用淤泥块代替, 具有较强的迷惑性. 为验证本方法在真实场景的攻击可复现性, 分别在实验室和真实环境中对车牌识别系统展开攻击, 并且将对抗样本用于开源的商业软件中进行测试, 验证了攻击的迁移性.

关键词 深度学习, 车牌识别, 对抗攻击, 黑盒攻击, 物理攻击

引用格式 陈晋音, 沈诗婧, 苏蒙蒙, 郑海斌, 熊晖. 车牌识别系统的黑盒对抗攻击. 自动化学报, 2021, **47**(1): 121−135

DOI 10.16383/j.aas.c190488

Black-box Adversarial Attack on License Plate Recognition System

CHEN Jin-Yin[1] SHEN Shi-Jing[1] SU Meng-Meng[1] ZHENG Hai-Bin[1] XIONG Hui[1]

Abstract Deep neural network (DNN) is one of the most commonly used deep learning methods and is widely used in various fields. However, DNN is vulnerable to adversarial attacks, so it is crucial to detect the vulnerabilities of DNN in the application system by adversarial attacks. In this paper, the vulnerability detection of the license plate recognition system is carried out. Under the premise of completely unknown internal structure information of the model, a black-box adversarial attack is launched, and security vulnerabilities in commercial license plate recognition system are found. The paper first proposes a black-box attack method for license plate recognition based on NSGA-II. Only by obtaining the output class label and corresponding confidence can produce a robust attack against environmental changes, and the algorithm controls the perturbation as a pure black block, which can be replaced by a silt block and has strong confusion. In order to verify the reproducibility of the attack of this method in real scenes, the license plate recognition system was attacked in the laboratory and the real environment, and the adversarial examples were tested in open source commercial software to verify the transferability of the attack.

Key words Deep learning, license plate recognition, adversarial attack, black-box attack, physical attack

Citation Chen Jin-Yin, Shen Shi-Jing, Su Meng-Meng, Zheng Hai-Bin, Xiong Hui. Black-box adversarial attack on license plate recognition system. *Acta Automatica Sinica*, 2021, **47**(1): 121−135

深度学习因其强大的特征提取和建模能力为人工智能的发展提供了巨大的机遇[1]. 其中, 深度神经网络 (Deep neural network, DNN) 作为最常用的深度学习方法之一, 在计算机视觉[2]、自然语言处理[3]、工业控制[4]、生物信息[5] 等众多研究领域获得成功. 同时 DNN 广泛应用于实际生活中, 如面部识别[6]、语音识别[7]、车牌识别等, 与我们的日常生活密不可分.

随着 DNN 的应用普及, 其安全问题也日益凸显, 已有研究表明 DNN 容易受到对抗样本的攻击[8], 即在正常样本上添加精心设计的微小对抗扰动后, DNN 将以较高的置信度输出错误类标. 更糟糕的是, 对抗样本可同时欺骗多种不同的 DNN 模型[9]. 根据攻击者是否已知目标模型的内部结构, 攻击可分为白盒攻击和黑盒攻击; 根据实施攻击的应用场景不同, 攻击可分为数字空间攻击和物理空间攻击[10]. 虽然数字空间中的白盒攻击产生的对抗扰动是人眼不可见的, 但是攻击时需要获取模型内部结构信息且对抗扰动难以精确打印, 因此难以应用于实际系统[11].

收稿日期 2019-07-01 录用日期 2019-12-23
Manuscript received July 1, 2019; accepted December 23, 2019
国家自然科学基金 (62072406), 浙江省自然科学基金 (LY19F020025), 宁波市 "科技创新 2025"重大专项 (2018B10063) 资助
Supported by National Natural Science Foundation of China (62072406), the Natural Science Foundation of Zhejiang Province (LY19F020025), the Major Special Funding for "Science and Technology Innovation 2025" of Ningbo (2018B10063)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 浙江工业大学信息工程学院 杭州 310023
1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023

基于 DNN 的车牌识别系统已广泛应用于生活, 如不停车收费系统 (ETC)、停车场自动收费管理、道路电子警察等. 本文首次提出了物理空间中车牌识别系统的黑盒攻击方法, 即在未知模型内部结构信息的前提下生成有效的对抗样本, 实现对商业车牌识别系统的物理攻击. 本文通过对抗攻击发现系统漏洞, 为进一步提高车牌识别系统的鲁棒性提供研究基础.

在物理空间中对车牌识别系统展开攻击, 除了需要考虑扰动尽可能小的限制条件外, 还需要克服以下几个困难: 1) 物理空间中的攻击成功率易受拍摄环境的影响 (如角度、距离、光线等); 2) 添加的扰动应具有迷惑性, 即人眼虽然能观测到, 但不认为是一种攻击手段; 3) 制作扰动时受现实条件制约, 例如打印出来的扰动存在色差, 导致对抗样本在物理空间中失效.

针对以上物理空间攻击存在的困难, 本文提出基于精英策略的非支配排序遗传算法 (NSGA-II)[12] 的黑盒攻击, 通过模型的输出类标及对应置信度生成对抗样本. 此外, 本文对样本进行角度、距离、光线等模拟变换, 并将变换后的攻击效果作为优化目标之一, 提高了对抗样本对环境因素的鲁棒性. 本文将扰动控制为纯黑色块, 可用车子行驶过程中飞溅的淤泥块代替. 扰动可被摄像头捕获导致系统错误识别, 但人将其视为正常污渍.

为验证本方法在物理环境中的可实现性, 分别在实验室和真实环境中对车牌识别系统展开攻击. 在实验室环境中, 通过改变角度、距离、光线等环境因素, 验证本文攻击具有较强的鲁棒性; 在真实环境中, 分别对三种车牌识别场景 (躲避公路上探头抓拍、躲避车牌尾号限行措施、冒充出入库车辆) 进行攻击. 此外, 本文还将对抗样本用于开源的商业软件中进行测试, 验证了攻击方法的迁移性.

综上所述, 本文的主要创新点包括:

1) 首次提出物理空间中车牌识别系统的黑盒攻击方法, 仅通过模型的输出类标和对应置信度即可造成有效攻击, 并成功攻击商用车牌识别系统.

2) 提出基于 NSGA-II 多目标优化的黑盒攻击方法, 在对抗样本生成过程中引入角度、距离、光线等环境因素, 提高了对抗样本的鲁棒性.

3) 提出多种真实场景中的车牌识别攻击方案, 实现了对商用车牌识别系统的漏洞检测.

1 相关工作

深度学习的对抗攻击是指在模型测试阶段, 攻击者通过在正常样本上添加精心设计的微小扰动得到对抗样本, 使得 DNN 误判的恶意攻击[10]. 已有的对抗攻击方法主要包括:

白盒攻击. Goodfellow 等[13] 提出快速梯度符号法 (Fast gradient sign method, FGSM), 通过梯度方向搜索快速产生对抗攻击效果的扰动. Kurakin 等[14] 提出基本迭代法, 采用小的搜索步长进行迭代计算扰动, 并将 BIM 扩展为迭代极小可能类法, 得到具有更强的对抗攻击性和较弱迁移性的对抗样本. Moosavi-Dezfooli 等[15] 提出 DeepFool 方法, 能够以更小的扰动实现与 FGSM 相近的攻击效果. Carlini 等[16] 提出 C&W 攻击, 通过限制不同的范数实现了较强的攻击效果. Papernot 等[17]提出了基于雅可比的显著图的攻击, 通过限制扰动范数 L_0, 得到扰动数量最少的对抗样本. Lyu 等[18]构建了对抗样本的正则化项, 并提出了三种基于梯度的攻击方法.

黑盒攻击. Su 等[19] 提出了单像素点攻击方法, 利用差分进化概念仅通过修改单个像素点即可生成对抗样本. Brendel 等[20] 提出了基于决策的边界攻击算法 Boundary, 从较大的对抗性扰动开始, 在保持对抗性的同时逐渐减少扰动. Chen 等[21] 提出了一种基于零阶优化的攻击方法 ZOO, 通过直接估计目标模型的梯度来生成对抗性的例子. Tu 等[22] 提出了 ZOO 的改进方法 AutoZOO, 利用基于自编码的零阶优化算法大大减少了对目标模型的访问次数. Chen 等[23] 提出了 Boundary++, 通过将 Boundary 与 ZOO 相结合的方法弥补了 Boundary 算法中模型访问次数过多的缺陷. Chen 等[24] 提出了基于进化计算的黑盒攻击方法, 实现有效的黑盒攻击. Bhagoji 等[25] 提出一种新的基于梯度估计的黑盒攻击方法, 该方法需要对目标模型的类概率进行查询访问, 并具有不依赖于攻击对象的迁移性.

物理空间的攻击. Kurakin 等[14] 首次证明了现实世界中攻击的存在. 在物理空间中的对抗攻击是指在现实环境中对系统进行攻击, 目标模型大多为落地服役模型, 攻击手段为现实可操作手段. 与上文介绍的数字空间中的攻击相比, 其主要困难在于误导实际应用系统的攻击者往往不能精确控制系统的输入. 相反, 攻击者可能只能控制它们的物理外观. 将物理场景转换为数字输入的过程不受攻击者的控制, 且转换过程易受光照、姿态和距离等因素的影响. 因此, 在实际操作方面, 物理攻击的攻击者更难以操纵或制造可能错误分类的输入. 基于深度学习模型的识别系统通常采用两步完成功能, 即第一步利用检测器定位待识别的对象, 第二步利用分类器识别定位到的对象. 根据攻击对象的不同, 物理攻击可分为针对检测器的物理攻击 (定位错

误) 和针对分类器的物理攻击 (识别错误).

1) 针对检测器的物理攻击: Chen 等[26] 提出了针对 Faster R-CNN[27] 的物理对抗攻击, 通过在停车标志上添加对抗扰动使 Faster R-CNN 错误检测. Eykholt 等[28] 提出了一种消失攻击, 通过在停车标志的红色区域内填充一个图案使得 YOLO-v2[29] 检测器无法检测到该停车标志, 并且该攻击可以转移到 Faster R-CNN. Thys 等[30] 提出了一种基于补丁的对抗攻击, 攻击者只需要在人身上放置一块补丁, 就能使检测器无法检测到此人.

2) 针对分类器的物理攻击: Sharif 等[31] 提出了针对人脸识别系统的面部攻击, 攻击者只需戴上一副特制的眼镜, 分类器就会作出攻击者预期的错误判断. Eykholt 等[32] 提出了针对路牌识别系统的物理攻击, 攻击者只需在路牌上添加一些不引人注意的涂鸦或者替换路牌背景, 路牌识别系统就会作出错误的判断. Sitawarin 等[33] 通过在现实环境中修改无害的标志和广告, 使它们被分类为攻击者所期望的具有高度置信度的交通标志, 该攻击扩大了对自驾车领域的威胁范围, 因为攻击者不只是局限于修改现有的交通标志. Athalye 等[34] 利用 3D 打印技术制作出了第一个物理对抗样本, 证明了三维对抗对象在物理世界中的存在. Li 等[35] 通过在相机镜头上放置一个精心制作的半透明贴纸, 使得拍摄到的图像具有对抗性, 并且镜头上的扰动为物理攻击中的通用扰动.

2　车牌识别系统的黑盒攻击方法

针对真实场景中的车牌识别系统, 本文提出了一种基于 NSGA-II 的黑盒攻击方法, 通过模型的输出类标及对应置信度生成对抗样本. 此外, 本文对样本进行角度、距离、光线等模拟变换, 并将变换后的攻击效果作为优化目标之一, 提高了对抗样本对环境因素的鲁棒性. 本文将扰动控制为纯黑色块, 可用车子行驶过程中飞溅的淤泥块代替. 扰动可被摄像头捕获导致系统错误识别, 但人将其视为正常污渍. 本文方法的整体框图如图 1 所示.

车牌识别系统的黑盒攻击方法主要步骤包括:

1) 对正常车牌从不同距离、角度拍摄一段视频, 截取其中若干帧作为正常车牌图像数据集;

2) 在正常车牌图像上分别添加随机扰动像素块, 生成多张不同的初始对抗样本, 构成初始种群;

3) 分别考虑攻击效果、扰动大小、环境影响等因素, 设计多目标优化函数, 并计算种群中每个样本的适应度值;

4) 对种群中的样本进行非支配排序和拥挤度排序, 选取一定数量的样本构成父代样本种群 P;

5) 判断是否达到迭代终止条件, 如果是, 执行步骤 8); 如果否, 执行步骤 6);

6) 对于父代种群 P, 使用交叉操作产生子代种群 Q;

7) 将父代种群 P 与子代种群 Q 合并为一个整

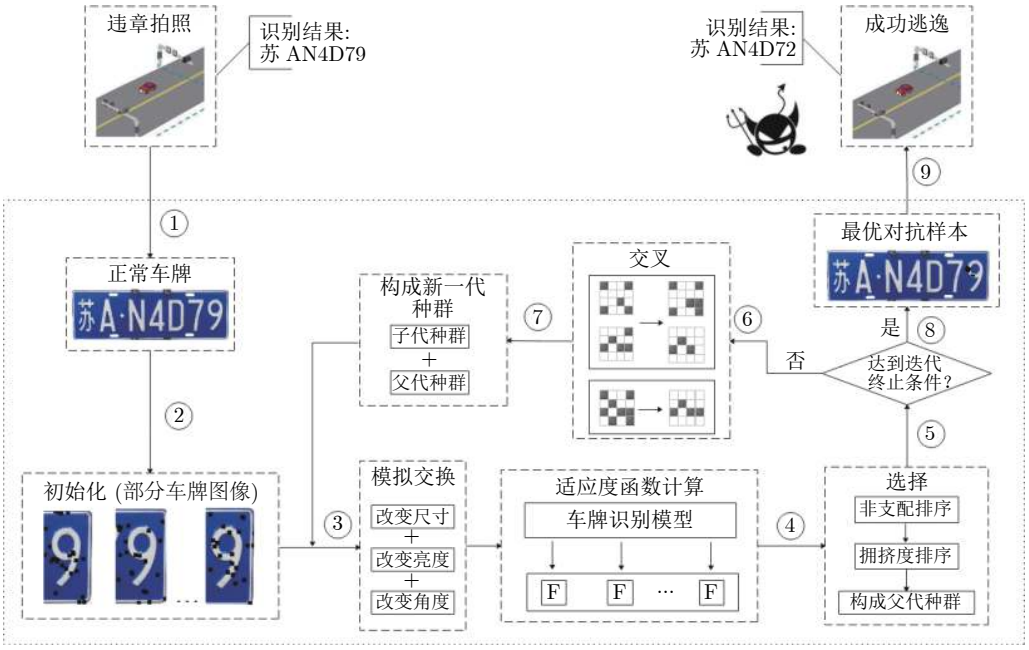


图 1　车牌识别系统的黑盒攻击方法整体框图

Fig. 1　The main block diagram of the proposed method against license plate recognition system

体种群 R, 跳转到步骤 3);

8) 选取符合要求的最优样本 (原类标置信度小于 0.2 且扰动最小);

9) 将扰动复现在真实车牌上, 实现物理攻击.

2.1 生成初始对抗样本

为产生黑盒攻击, 在正常车牌数据集中选取一张车牌图像样本 x, 在 x 上添加随机扰动块, 构成初始对抗样本. 为保证生成的扰动可在实际车牌识别系统中有效, 本文利用扰动像素块代替一般图像对抗样本中的单个像素的扰动, 即每张车牌图像上添加若干个一定大小的初始黑色扰动块, 每个扰动块的大小可根据车牌所占像素大小进行调节, 初始扰动块的参数敏感性分析可见第 3.6.1 节. 扰动块位置随机分布, 为了在物理空间中准确复刻对抗扰动, 本文将扰动设置为块状且为纯黑色.

2.2 基于多目标优化的适应度函数设计

为实现车牌识别系统的成功攻击, 对抗样本需要满足两个优化目标: 添加的对抗扰动受限、识别类标错误. 本文提出多目标优化的适应度函数设计.

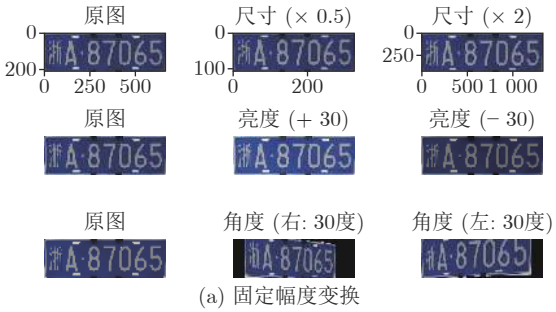
对抗扰动受限. 计算每张车牌图像样本中扰动总数的面积, 即所有扰动的像素点个数总和, 记为 $S = \|x' - x\|_0$, 其中 x' 表示车牌对抗样本, x 表示车牌正常样本. 将 S 作为优化目标 F_1, F_1 越小, 则表示对抗样本的添加扰动越小.

识别类标错误. 为了实现车牌识别系统的错误识别, 本文引入目标函数 $\min f(x')_y$, 其中 $f(x')_y$ 表示对抗样本 x' 被分为第 y 类的置信度, y 表示正常样本 x 的正确分类结果. $f(x')_y$ 越小, 表示车牌对抗样本 x' 被错误分类的概率越高.

为提高车牌对抗样本在物理场景中的攻击成功率, 克服环境因素变换对攻击效果的影响, 本文利用图像处理技术模拟真实拍摄场景下的三种变换, 并将变换后的分类结果加入优化目标. 三种变换分别是: 利用图像缩放变换模拟不同拍摄距离; 利用图像亮度变换模拟不同拍摄光线; 利用图像透视变换模拟不同拍摄角度. 三种变换的幅度选取方式分为固定幅度和随机幅度两种.

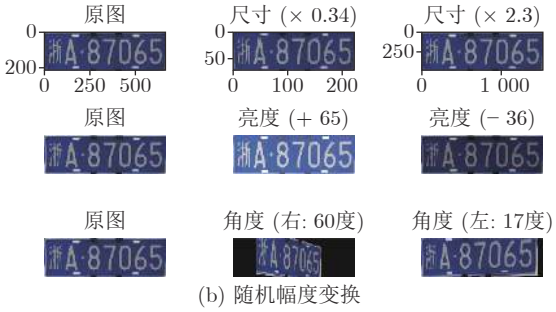
图 2 列举了两种幅度选取方式的变换效果图. 图 2 (a) 展示了固定幅度时的变换效果图. 第一行模拟了距离变换, 分别将图像尺寸 (长宽) 缩小至 0.5 倍, 放大至 2 倍; 第二行模拟了亮度变换, 分别将图像像素值增大 30, 减小 30; 第三行模拟了角度变换, 分别向右倾 30 度, 向左倾 30 度.

图 2 (b) 展示了随机幅度时的变换效果图. 第一行模拟了距离变换, 分别将图像尺寸 (长宽) 随机缩小至 0.34 倍, 放大至 2.3 倍; 第二行模拟了亮度变换, 分别将图像像素值随机增大 65, 减小 36; 第三行模拟了角度变换, 分别向右随机倾斜 60 度, 向左随机倾斜 17 度.



(a) 固定幅度变换

(a) Fixed amplitude transformation



(b) 随机幅度变换

(b) Random amplitude transformation

图 2 模拟场景变换效果图

Fig. 2 Scene simulation change effect diagram

将模拟变换后样本的识别结果作为优化目标的一部分, 则设计目标函数 F_2 的计算公式如下:

$$F_2 = f(x')_y + \frac{\sum\limits_{i=1}^{Num} f(T_i(x'))_y}{Num} \tag{1}$$

其中, Num 表示图像变换种类; $T_i(x')$ 表示变换形态后的对抗样本, $i = 1, 2, \cdots, Num$; $f(T_i(x'))_y$ 表示变换后的样本被分为第 y 类的置信度; y 表示正常样本 x 的正确分类结果. F_2 越小, 则表示攻击效果越好.

以图 3 的两张车牌对抗样本为例, 两张样本均被识别为 “浙 AS7065”. 经计算, 样本 (a) 的 $F_1 = 156.0$, $F_2 = 0.197$; 样本 (b) 的 $F_1 = 237.0$, $F_2 =$

(a) 浙 AS7065
(a) Zhe AS7065

(b) 浙 AS7065
(b) Zhe AS7065

图 3 车牌对抗样本

Fig. 3 License plate adversarial examples

0.015. 两张样本对原车牌 “浙 A87065”均攻击成功, 但是样本 (a) 的扰动小于样本 (b), 样本 (b) 的攻击鲁棒性强于样本 (a).

2.3　选择算子

本文采取基于 NSGA-II 的算法构成父代样本种群. 首先将子代车牌对抗样本与父代车牌对抗样本合并为一个种群 R, 然后分别进行非支配排序和拥挤度排序, 选出前 N 个优秀对抗样本作为父代个体进行第 2.4 节的交叉操作.

非支配排序. 当车牌样本 x'_A 中所有目标都优于或等于车牌样本 x'_B 时 (即 x'_A 中的 F_1 和 F_2 均小于等于 x'_B 中的 F_1 和 F_2), 则定义为 x'_A 支配了 x'_B, 否则 x'_A 和 x'_B 是一种非支配关系. 非支配排序的整体思路是对种群 R 中的车牌样本进行等级划分, 即分为 $Rank_0$, $Rank_1$, $Rank_2, \cdots$. 其中, $Rank_0$ 中的车牌样本均优于 $Rank_1$ 中的车牌样本, 以此类推. 快速非支配排序算法如下:

1) 计算每张车牌样本 x'_i 的两个参数 n_i 和 s_i, 其中 n_i 表示种群 R 中支配 x'_i 的车牌样本数目, s_i 表示种群 R 中被 x'_i 支配的车牌样本集合;

2) $k = 0$, k 表示划分的等级;

3) 寻找种群 R 中所有 $n_i = 0$ 的车牌样本 x'_i, 保存在等级 $Rank_k$ 中;

4) 对于 $Rank_k$ 中的每张车牌样本 x'_i, 遍历 s_i 中的每张车牌样本 x'_l, 执行 $n_l = n_l - 1$, 若 n_l 等于 0, 则将 x'_l 保存在集合 $Rank_{k+1}$ 中;

5) $k = k + 1$, 进入下一等级的划分;

6) 重复步骤 4) 和步骤 5), 直到整个种群 R 被划分完毕.

拥挤度排序. 为了判定同一个 $Rank$ 层中车牌样本的优劣, 将每张车牌样本的拥挤度作为评价标准. 拥挤度越大表示该车牌样本与其他样本之间的差异性越大, 也意味着该车牌样本越优. 拥挤度排序用于保持每代车牌样本的多样性. 每个样本的拥挤度计算方式如下:

$$i_d = \sum_{j=1}^{m} (|f_j^{i+1} - f_j^{i-1}|) \tag{2}$$

其中, i_d 表示第 i 个车牌样本的拥挤度, m 表示有 m 个目标函数 (本文 $m = 2$, 即扰动总面积与原类标置信度), f_j^{i+1} 表示第 $i+1$ 个车牌样本的第 j 个目标函数值.

非支配排序与拥挤度排序计算完毕, 进入车牌样本选择过程. 设定每次迭代种群中需要选择的车牌样本数量为 N, 每次挑选时, 先挑选表现最好的样本, 即 $Rank_0$ 中的车牌样本, 接着 $Rank_1$, $Rank_2$, $Rank_3, \cdots$. 但是总会出现以下情况: $\sum_{k=0}^{n-1} Rank_k < N$ 且 $\sum_{k=0}^{n} Rank_k > N$, 此时需要通过拥挤度排序选出 $Rank_n$ 层中较优的车牌样本, 即计算 $Rank_n$ 层中每张车牌样本的拥挤度, 再根据拥挤度从大到小排序, 选出拥挤度大的车牌样本. 最终两种排序方式选出的个体总数为 N, 构成下一次迭代的父代种群.

本文以车牌 “浙 A87065”排序过程为例, 说明选择算子的具体操作过程. 图 4 为车牌 “浙 A87065” 第 10 次攻击迭代中子代加父代种群的非支配排序结果. 在进行非支配排序时, 为节省时间成本, 不需要将种群全部划分, 本次排序只划分到 $Rank_3$, 因为 $\sum_{k=0}^{2} Rank_k < N$ 且 $\sum_{k=0}^{3} Rank_k > N$, 所以需要先计算 $Rank_3$ 中车牌样本的拥挤度 (如图 4, 样本 x' 的拥挤度为 $d_1 + d_2$); 然后进行拥挤度排序, 选出 $Rank_3$ 中拥挤度大的样本, 将这些样本与 $Rank_0$、$Rank_1$、$Rank_2$ 中的样本共同组成下一代父代种群.

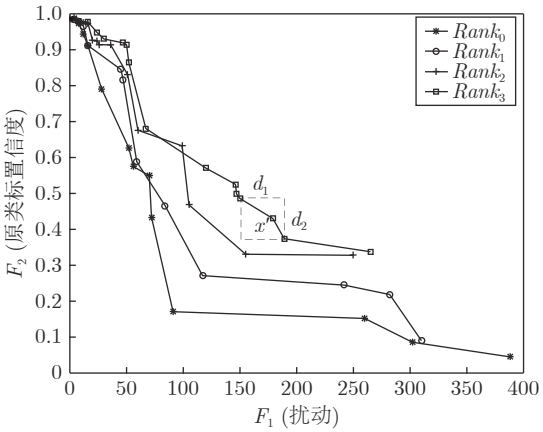


图 4　车牌样本第 10 次迭代非支配排序结果

Fig. 4　License plate example in 10th iteration non-dominated sorting result

2.4　交叉算子

本文采用随机交叉算法生成新的子代个体. 从父代车牌样本种群中随机选取两个样本, 记录两个样本的扰动块数量分别为 n_1 和 n_2, 从两个样本中随机选取 $a \in (0, n_1)$ 和 $b \in (0, n_2)$ 个扰动块, 将 a 和 b 个扰动合成一个子代, 余下的 $n_1 - a$ 和 $n_2 - b$ 个扰动合成另一个子代. 随机交叉过程示意图如图 5 所示, 其中黑色区域表示扰动块. 图 6 表示车牌样本的交叉示例, 黑点小块表示扰动.

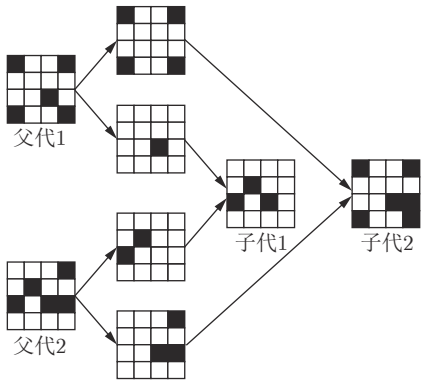


图 5 随机交叉过程示意图

Fig. 5 Schematic diagram of the random cross process

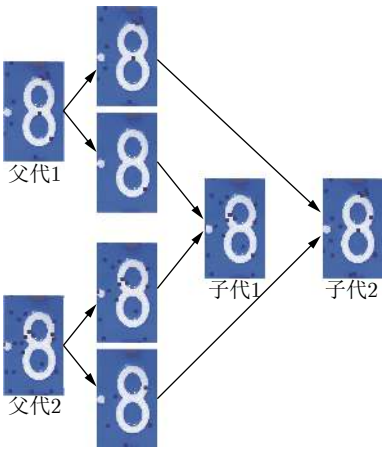


图 6 车牌样本随机交叉示例

Fig. 6 Example of random intersection of license plate examples

2.5 最优样本获取

当达到最大迭代次数时, 从最后一代种群的 $Rank_0$ 中选取最优对抗样本. 最优对抗样本的选取可根据攻击情景决定, 对扰动隐蔽性要求高的可选取攻击成功中扰动较小的对抗样本; 对样本鲁棒性要求高的可选取原类标置信度较小的对抗样本. 如图 7 表示对车牌 "浙 A87065"中数字 "8" 攻击结果的帕累托曲线, 其中 k 表示第 k 次迭代, 每条曲线表示该次迭代中的最优样本群体 ($Rank_0$). 本文将最优对抗样本定义为原类标置信度小于 0.2 时 (以较高的置信度攻击成功) 扰动最小的样本, 即图中 "#" 所指的位置.

3 实验与分析

本节内容分别从数据集与识别模型、评价指标、车牌图像攻击算法对比、环境模拟变换中的车牌图像对抗样本、实验室环境的车牌识别系统攻击、参数敏感性分析、现实场景中商用车牌识别系统的攻击等方面展开.

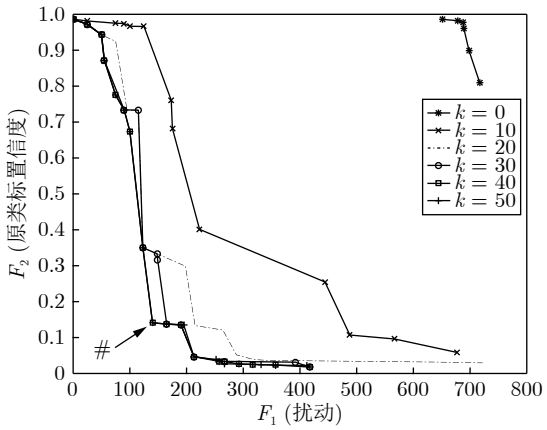


图 7 数字 "8"攻击结果的帕累托曲线

Fig. 7 Pareto curve of the number "8" attack result

3.1 数据集与识别模型

CCPD[36] 是国内用于车牌识别的大型数据集, 由中科大构建, 该数据集绝大多数为 "皖 A"车牌. 本文为扩充完善各个省份的车牌, 将另一公开数据集[1] 中的车牌数据与 CCPD 组合, 作为本文的车牌数据集, 共 5 000 张车牌样本. 此外, 本文又在 5 000 张车牌样本上作了旋转、模糊和亮度调节等图片数据增强操作, 再次扩充了数据集的数量, 达到 20 000 张图片数据集.

现阶段应用较广的中国车牌识别开源模型包括三种: HyperLPR[2]、EasyPR[3]、基于 PaddlePaddle 的 OCR 车牌识别[4]. 为比较三者的有效性, 本文利用上述公开车牌数据集分别对三种模型进行了训练验证检测, 并统计其识别准确率, 如表 1 所示. (注: HyperLPR 存在预训练模型, 本文主要对预训练模型进行微调.)

表 1 三种模型的识别准确率

Table 1 Recognition accuracy of the three models

模型名称	训练准确率	测试准确率
HyperLPR	96.7%	96.3%
EasyPR	85.4%	84.1%
PaddlePaddle - OCR车牌识别	87.9%	86.5%

车牌识别对比实验设置为 1) 数据集: 公开数据集 20 000 张车牌图像; 2) 实验平台及环境: i7-

[1] https://pan.baidu.com/s/1RyoMbHtLUlsMDsvLBCLZ2w
[2] https://github.com/zeusees/HyperLPR
[3] https://github.com/liuruoze/EasyPR
[4] https://github.com/huxiaoman7/mxnet-cnn-plate-recognition

7700K 4.20GHzx8 (CPU), TITAN Xp 12GiBx2 (GPU), 16GBx4 memory (DDR4), Ubuntu 16.04 (OS), Python 3.6, Tensorflow-gpu-1.3.

由表 1 可知, HyperLPR 模型的识别准确率远高于其他两种模型, 因此本文将 HyperLPR 模型作为本文攻击算法的目标模型, 验证本文攻击方法的有效性.

此外, 本文将 HyperLPR 模型生成的对抗样本对百度 AI 开放平台中的车牌识别板块[5] 及商业软件 OpenALPR[6] 进行测试, 验证了本文提出的黑盒攻击具有较强的攻击迁移性.

3.2　评价指标

本文利用攻击成功率 (Attack success rate, ASR)、扰动大小、样本鲁棒性 (原类标置信度) 和收敛时的迭代次数来评价实验的性能.

攻击成功率计算方式如式 (3) 所示:

$$\mathrm{ASR}=\frac{sumNum(label(x')\neq y_0)}{sumNum(label(x)=y_0)} \tag{3}$$

其中, $sumNum(\cdot)$ 表示样本数量, x 表示原图, x' 表示对抗样本, $label(\cdot)$ 表示输出的类标, y_0 表示正常车牌的正确类标.

扰动大小用来评价对抗样本中扰动的隐蔽性. 本文采用平均 L_0 范数 ($\tilde{L}_0$) 和平均 L_2 范数 ($\tilde{L}_2$) 两种扰动计算方式. 本文算法生成的对抗样本中的扰动为纯黑色, 重点关注扰动的区域面积, 所以 $\tilde{L}_0$ 范数是本文衡量扰动大小的主要指标. $\tilde{L}_0$ 和 $\tilde{L}_2$ 的计算方式如式 (4) 和式 (5) 所示:

$$\tilde{L}_0=\frac{\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}\mathrm{sign}(x'_{ijk}-x_{ijk})}{h\times w\times c} \tag{4}$$

$$\tilde{L}_2=\frac{\sqrt{\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}(x'_{ijk}-x_{ijk})^2}}{h\times w\times c} \tag{5}$$

其中, h、w、c 分别表示图像中的高度、宽度、通道数, x'_{ijk} 表示对抗样本在第 k 个通道上坐标为 $(i,\ j)$ 的像素点的值, x_{ijk} 表示原始图像在第 k 个通道上坐标为 $(i,\ j)$ 的像素点的值, 图像的像素值范围是 0 到 255.

样本鲁棒性用原类标置信度表示, 对抗样本被识别成原类标的置信度越低, 表示该样本越鲁棒.

收敛时的迭代次数主要用来衡量一次攻击所需要的时间成本和访问代价. 本文算法对模型的访问次数 = 迭代次数 × 种群大小.

[5] http://ai.baidu.com/tech/ocr/plate

[6] http://www.openalpr.com/cloud-api.html

3.3　车牌图像攻击算法对比

在基于深度学习的计算机视觉领域, 现已有多种对抗攻击的算法. 其中, 白盒攻击中较为典型的有 FGSM 和基于 2-norm 的攻击方法, 黑盒攻击中较为典型的有 ZOO 和 AutoZOO 攻击方法. 本文针对车牌识别系统提出一种基于 NSGA-II 的黑盒攻击方法, 实现了物理空间的对抗攻击.

NSGA-II 算法参数设置如下: 种群规模为 50 张车牌样本, 初始扰动所占总面积比为 1: 80, 扰动块数量为 30, 形状为矩形, 进化停止代数为 50 代, 交叉概率为 0.8.

本文面向车牌数据集, 分别采用 4 种性能较优的白盒与黑盒攻击算法对每张图像上的 7 个字符展开对抗攻击, 目标模型为 HyperLPR, 并将攻击结果同本文算法的攻击结果进行对比, 如表 2 所示. 为了配合 FGSM 等主流攻击算法中的扰动计算方式, 表 2 中的扰动大小用平均 L_2 范数 ($\tilde{L}_2$) 和平均 L_0 范数 ($\tilde{L}_0$) 计算.

表 2　车牌图像攻击算法对比结果

Table 2　Comparison results of plate images attack algorithms

攻击算法		攻击成功率	$\tilde{L}_2$	$\tilde{L}_0$	访问次数
白盒	FGSM	89.3%	0.067	0.937	32
	2-norm	92.8%	0.051	0.923	3
黑盒	ZOO	85.7%	0.087	0.953	74356
	AutoZOO	87.1%	0.069	0.938	4256
	本文算法	98.6%	0.035	0.004	1743

由表 2 可知, 在指定攻击车牌上的某一个字符时, 本文算法的攻击成功率高于其他 4 种攻击方法, 并且 $\tilde{L}_2$ 扰动也相应更小一点, 其中一张车牌的对抗样本如图 8 所示. $\tilde{L}_0$ 扰动主要衡量扰动的像素点个数, 即扰动区域. 由于本文算法重点限制扰动区域, 而其他 4 种攻击方法在整张车牌上添加扰动, 所以本文算法的 $\tilde{L}_0$ 扰动小于其他 4 种攻击方法. 就黑盒攻击而言, 本文算法对模型的访问次数远小于其他两种攻击算法, 这得益于本文算法只需较少的迭代次数. 此外, 白盒攻击对模型的访问次数远小于黑盒攻击对模型的访问次数, 基于 2-norm 的攻击方法在各项指标上均优于 FGSM, AutoZOO 在各项指标上也均优于 ZOO.

由图 8 的对抗样本生成结果可知, 前 4 种攻击算法得到的对抗扰动基本都是全局分散的, 而且每个扰动的像素值几乎都不相同, 有的扰动甚至肉眼无法观测到. 很显然前 4 种攻击方法得到的车牌对抗样本难以在物理空间中复现, 即使通过打印纸质车牌, 依旧存在打印色差的问题. 但是本文算法只

(a) FGSM对抗样本图. 图像从左到右依次表示攻击车牌第一个字符、第二个字符, 以此类推
(a) Adversarial examples of FGSM. The image from left to right represents attacking the first character, the second character, and so on

(b) 2-norm对抗样本图. 图像从左到右依次表示攻击车牌第一个字符、第二个字符, 以此类推
(b) Adversarial examples of 2-norm. The image from left to right represents attacking the first character, the second character, and so on

(c) ZOO对抗样本图. 图像从左到右依次表示攻击车牌第一个字符、第二个字符, 以此类推
(c) Adversarial examples of ZOO. The image from left to right represents attacking the first character, the second character, and so on

(d) AutoZOO对抗样本图. 图像从左到右依次表示攻击车牌第一个字符、第二个字符, 以此类推
(d) Adversarial examples of AutoZOO. The image from left to right represents attacking the first character, the second character, and so on

(e) 本文算法对抗样本图. 图像从左到右依次表示攻击车牌第一个字符、第二个字符, 以此类推
(e) Adversarial examples of ours. The image from left to right represents attacking the first character, the second character, and so on

图 8 不同攻击算法攻击同一张车牌样本的不同位置的对抗样本图

Fig. 8 Adversarial examples graph of different attack algorithms attacking different positions of the same license plate sample

需要在真实车牌的相同位置贴上相同形状的黑色小纸片即可完成复现, 甚至粘上黑色泥土也能达到一定的攻击效果.

3.4 环境模拟变换中的车牌图像对抗样本

为提高对抗样本在物理攻击中的成功率, 本文利用 NSGA-II 自动生成数字 “0”到 “9”的对抗样本, 并在不同数字模拟环境中测试鲁棒性. 本文采取三套不同的幅度变换策略生成对抗样本, 分别是固定 1、固定 2、随机变换.

1) 固定 1: 将原始对抗样本尺寸缩小至 0.5 倍和放大至 2 倍; 将原始对抗样本像素值增加 30 和减小 30; 将原始对抗样本向右倾斜 30 度和向左倾斜 30 度.

2) 固定 2: 将原始对抗样本尺寸缩小至 0.3 倍和放大至 3 倍; 将原始对抗样本像素值增加 50 和减小 50; 将原始对抗样本向右倾斜 50 度和向左倾斜 50 度.

3) 随机变换: 将原始对抗样本尺寸随机缩小至 $S_1 \in (0.2,\ 1)$ 倍和放大至 $S_2 \in (1,\ 5)$ 倍; 将原始对抗样本像素值随机增加 $P_1 \in (0,\ 100)$和减小 $P_2 \in (0,\ 100)$; 将原始对抗样本随机向右倾斜 $A_1 \in (0,\ 60)$ 度和向左倾斜 $A_2 \in (0,\ 60)$ 度.

本文使用的车牌图像大小约为 350 × 100 个像素点, 当尺寸缩小倍数小于 0.2 时, 会对图像的可视性造成较大的影响; 当尺寸放大倍数大于 5 时, 会导致模型识别准确率的下降; 改变的像素值超过 $(-100,\ 100)$ 这个范围时, 会导致图像过暗或过亮, 降低模型的识别准确率; 倾斜角的阈值设定与文献 [32] 一致, 超过该阈值后, 模型的识别准确率会大大降低.

实验结果如表 3 所示. 其中, 左边第 1 列表示对对抗样本进行不同的环境模拟变换; 左边第 2 列表示三套不同的幅度变换策略; 第 3 列到第 12 列具体列举了数字 “0”到 “9”的攻击成功率, 每一个数字都生成了三种不同环境模拟策略的对抗样本, 所以每一个数字对应的每一种环境下, 都有三个攻击成功率, 分别是策略 “固定 1”、“固定 2”和 “随机变换”. 原始对抗样本表示不对样本作模拟变换, 对于这种情况, “固定 1”策略的攻击成功率最高, 这得益于 “固定 1”策略中环境变换幅度较小; “尺寸 (× 0.5)、光线 (+ 30)、角度 (右 30 度)”和 “尺寸 (× 2)、光线 (– 30)、角度 (左 30 度)”这两种环境模拟变换与 “固定 1”策略中的变换相同, 所以由 “固定 1”生成的对抗样本的攻击成功率高于其他两种; “尺寸 (× 0.3)、光线 (+ 50)、角度 (右 50 度)”和 “尺寸 (× 3)、光线 (– 50)、角度 (左 50 度)”这两种环境模拟变换与 “固定 2”策略中的变换相同, 所以由 “固定 2”生成的对抗样本的攻击成功率高于其他两种, 并且在这种情况下 “随机变换”策略优于 “固定 1”;

表 3　不同环境模拟策略在不同模拟环境下的攻击成功率

Table 3　The attack success rate of different simulation strategies in different simulation environments

环境因素		攻击成功率 (%)										平均成功率 (%)
		0	1	2	3	4	5	6	7	8	9	
原始对抗样本	固定1	100	96	100	100	100	100	100	100	100	100	**99.6**
	固定2	100	94	96	98	94	100	100	96	100	100	98.0
	随机变换	100	94	98	100	94	100	100	98	100	100	98.4
尺寸 (× 0.5)	固定1	100	80	90	92	90	94	98	90	96	100	**93.8**
光线 (+ 30)	固定2	98	76	90	84	88	92	94	86	92	98	90.4
角度 (右 30 度)	随机变换	100	76	92	92	90	94	96	88	92	98	92.8
尺寸 (× 2)	固定1	100	80	90	92	92	94	98	90	96	100	**93.6**
光线 (– 30)	固定2	100	78	90	86	86	88	92	82	90	96	89.2
角度 (左 30 度)	随机变换	100	78	92	90	88	90	96	84	92	96	91.4
尺寸 (× 0.3)	固定1	92	76	80	86	82	84	88	84	90	88	85.0
光线 (+ 50)	固定2	98	82	92	92	90	94	96	90	96	98	**93.4**
角度 (右 50 度)	随机变换	96	80	90	88	86	90	94	92	92	94	90.8
尺寸 (× 3)	固定1	90	74	80	86	82	82	90	82	88	88	84.2
光线 (– 50)	固定2	98	80	90	92	90	92	96	92	94	98	**92.8**
角度 (左 50 度)	随机变换	96	78	88	88	84	92	92	92	94	94	90.6
尺寸 (× 0.7)	固定1	92	76	80	88	84	82	90	82	92	90	85.6
光线 (+ 20)	固定2	94	76	86	90	86	84	92	86	90	92	88.4
角度 (右 42 度)	随机变换	96	78	90	92	90	92	94	90	94	96	**92.2**
尺寸 (× 1.3)	固定1	92	76	78	86	82	84	90	82	88	84	84.2
光线 (– 75)	固定2	92	74	82	86	82	86	92	88	90	90	88.0
角度 (左 15 度)	随机变换	94	76	88	90	86	92	92	90	92	94	**91.0**
各种环境平均攻击成功率	固定1	95.1	**79.7**	85.4	90.0	87.4	88.6	93.4	87.1	92.9	92.9	89.3
	固定2	97.1	79.4	89.4	89.4	88.0	90.9	94.6	88.3	93.1	**96.0**	90.8
	随机变换	**97.4**	**79.7**	**90.6**	**90.6**	**88.3**	**92.6**	**94.9**	**90.0**	**93.7**	**96.0**	**91.6**

“尺寸 (× 0.7)、光线 (+ 20)、角度 (右 42 度)”和 “尺寸 (× 1.3)、光线 (– 75)、角度 (左 15 度)”这两种环境模拟变换是随机选取的, 在这种情况下, “随机变换”策略优于前两种, 并且 “固定 2”优于 “固定 1”. 由此可知, 当环境模拟变换与进化时的模拟策略一致时, 固定幅度的攻击成功率会相应提升, 并且固定的幅度越大, 对外界环境的适应能力越好, 但是生成对抗样本时的成功率会略微下降. 当外界环境变换不可知时, “随机变换”策略生成的对抗样本的攻击鲁棒性更高, 由此可知随机幅度变换更适用于本文算法.

由表 3 的 “各种环境平均攻击成功率”可知, 不同数字的对抗样本的鲁棒性并不相同. 在不同的环境变换下, 数字 “0” 的三种对抗样本的攻击成功率均为最高, 数字 “1” 的三种对抗样本的攻击成功率均为最低. 换言之, 在物理攻击中数字 “0” 比数字 “1” 更容易被攻击.

表 4 展现了表 3 中数字 “0”到 “9”每类对抗样本中的其中一张样本的分类结果. 由表 4 可知, 同一张对抗样本在进行不同的环境模拟变换后, 错误分类的结果会不全相同, 这意味着本文算法在车辆逃逸场景中有更高的适用性. 此外, 三种策略下的数字 “1”均有被正确分类的情况, 该结果也进一步验证了本文算法中数字 “1”较难被攻击的结论, 其原因是本文算法在车牌上添加的扰动为黑色扰动块, 数字 “1”对黑色扰动块的敏感性较低, 与之相反, 对白色扰动块的敏感性较高. 最后, 无论是哪种策略的对抗样本, 在经过不同的变换后, 其识别置信度均有一定程度的下降, 且下降幅度的趋势与变换幅度呈正相关. 这一结果也从侧面反映了在现实场景中, 模型的识别结果易受到环境因素的影响.

3.5　实验室环境的车牌识别系统攻击

完成数字空间中的车牌图像攻击后, 将车牌对抗样本按车牌真实比例放大后打印, 然后将扰动裁剪下来, 按对应位置粘贴在真实车牌上, 测试本文

算法在实验室环境中的物理攻击效果, 调整实验室环境的距离、光线、角度, 检测对抗样本的鲁棒性. 实验结果如表 5 所示. 本实验的车牌识别模型是 HyperLPR, 原始正常车牌为 “苏 AN4D79”.

由表 5 可知, 正常车牌在不同物理环境下均能以 0.9 以上的置信度被正确识别. 之后, 我们分别在 “N”、“D”、“9” 上添加扰动 (粘上打印后裁剪下来的黑色纸张), 实验结果可得添加扰动后的车牌在不同的物理环境下均被错误识别为 “苏 AH4072”, 分类置信度有所下降, 但均高于 0.8, 属于正常的置信度范畴. 实验验证了数字空间中攻击成功的对抗样本在物理空间中复现后, 也具有较强的攻击能力.

3.6 参数敏感性分析

本文使用基于精英策略的非支配排序遗传算法 (NSGA-II), 主要的参数包括: 初始扰动信息、交叉概率、迭代次数. 在本节中, 对以上参数逐个进行敏感性分析.

3.6.1 初始扰动信息分析

本文将遗传进化算法作为主要的攻击方法, 初始种群 (初始扰动) 的信息对本文算法的有效性具有直接影响. 本文将初始扰动设定为若干个扰动块, 每个扰动块由若干个黑色像素点组成, 每个扰动块面积相等.

以下对初始扰动块所占面积比值、数量、形状进行具体分析. 初始扰动块所占面积比值表示车牌样本中所有扰动块面积之和与车牌面积之比; 扰动的数量表示每张车牌样本中扰动块的数目; 初始扰动的形状被设定为矩形 (R)、圆形 (C) 以及混合方块圆形 (R+C).

表 4 车牌对抗样本识别结果及其置信度、扰动等展示

Table 4 License plate against sample identification results and their confidence, disturbance display

环境因素	识别结果 (固定1/固定2/随机变换)									
										
原始对抗样本	C/C/Q	H/5/5	Z/Z/Z	5/5/2	J/6/Z	3/3/3	3/5/5	T/T/Z	G/S/S	2/2/2
	平均置信度: 0.92/0.87/0.86									
尺寸 (× 0.5) 光线 (+ 30) 角度 (右 30 度)	C/C/Q	H/5/5	Z/Z/3	5/5/2	J/X/Z	3/3/3	3/3/3	T/1/Z	G/S/S	2/2/2
	平均置信度: 0.90/0.86/0.83									
尺寸 (× 2) 光线 (– 30) 角度 (左 30 度)	C/C/Q	H/7/5	Z/Z/Z	5/5/2	J/6/Z	3/3/3	3/5/5	T/T/Z	G/S/G	2/2/2
	平均置信度: 0.89/0.83/0.86									
尺寸 (× 0.3) 光线 (+ 50) 角度 (右 50 度)	C/C/Q	1/5/1	2/Z/X	5/5/5	4/6/X	3/3/3	3/5/5	T/T/Z	G/S/S	2/2/2
	平均置信度: 0.84/0.90/0.85									
尺寸 (× 3) 光线 (– 50) 角度 (左 50 度)	C/C/Q	1/5/7	Z/Z/Z	5/5/2	J/6/4	3/3/3	3/5/5	1/T/1	G/S/S	2/2/2
	平均置信度: 0.84/0.88/0.86									
尺寸 (× 0.7) 光线 (+ 20) 角度 (右 42 度)	C/C/Q	H/1/5	Z/Z/Z	5/5/2	J/6/Z	3/3/3	5/5/5	T/T/Z	G/S/S	2/2/2
	平均置信度: 0.81/0.87/0.85									
尺寸 (× 1.3) 光线 (– 75) 角度 (左 15 度)	C/C/0	1/1/7	Z/2/Z	5/5/5	4/6/Z	3/3/3	3/5/5	7/T/Z	S/G/S	2/2/2
	平均置信度: 0.87/0.82/0.83									

表 5 实验室环境的车牌对抗攻击

Table 5 License plate adversarial attack in the laboratory environment

环境因素	0度, 1 m, 白天	0度, 1 m, 夜晚	0度, 5 m, 白天	0度, 5 m, 夜晚	20度, 1 m, 白天	20度, 1 m, 夜晚
物理对抗样本						
正常车牌识别结果	苏 AN4D79	苏 AN4D79	苏 AN4D79	苏 AN4D79	苏 AN4D79	苏 AN4D79
正常车牌识别置信度	0.9751	0.9741	0.9242	0.9214	0.9578	0.9501
对抗样本识别结果	苏 AH4072	苏 AH4072	苏 AH4072	苏 AH4072	苏 AH4072	苏 AH4072
对抗样本识别置信度	0.9041	0.8862	0.8248	0.8310	0.8045	0.8424

本文选取数据集中 10 张车牌样本, 分别用不同的初始扰动信息对车牌上的 7 个字符进行攻击, 一共生成 70 张对抗样本, 统计不同的初始扰动信息对实验结果的影响. 具体实验结果如表 6 所示, 其中 “最终扰动”列表示攻击得到的最优对抗样本的扰动大小 (用平均 L_0 范数计算).

表 6　初始扰动信息的影响

Table 6　Influences of initial perturbation information

面积比值	数量	形状	攻击成功率	最终扰动	迭代次数
1: 50	10	R	100%	0.0062	33
		C	100%	0.0059	32
		R+C	100%	0.0063	35
	30	R	100%	0.0054	36
		C	100%	0.0052	35
		R+C	100%	0.0054	34
	50	R	100%	0.0042	42
		C	100%	0.0043	40
		R+C	100%	0.0043	44
1: 80	10	R	100%	0.0058	34
		C	100%	0.0054	33
		R+C	100%	0.0055	34
	30	R	100%	0.0043	34
		C	100%	0.0041	32
		R+C	100%	0.0042	35
	50	R	96%	0.0037	48
		C	94%	0.0036	48
		R+C	96%	0.0032	46
1: 120	10	R	100%	0.0042	32
		C	100%	0.0045	31
		R+C	98%	0.0042	31
	30	R	98%	0.0035	36
		C	96%	0.0033	36
		R+C	96%	0.0033	35
	50	R	87%	0.0027	56
		C	86%	0.0025	58
		R+C	87%	0.0024	58

由表 6 可知, 在初始扰动块所占面积比值及数量一定的情况下, 初始扰动块的形状对本文算法攻击效果的影响可以被忽略. 在初始扰动块所占面积比值一定时, 扰动块的数量越多, 最终得到的对抗样本的扰动越小, 但是迭代次数会有所上升. 当初始扰动块的数量一定时, 扰动块所占面积比值越小, 最终扰动相应也会越小, 但是攻击成功率会相应地下降, 迭代次数也会上升.

不同字符的攻击难易程度也不同, 如汉字和一些数字 (“0”、“8”等) 更容易受到攻击, 此时我们可以采用面积比值较小的初始扰动来进行攻击. 但是像 “A” 这种字母, 相对更难攻击, 所以我们应该选择面积比值较大的初始扰动来进行攻击. 总之, 初始扰动块所占面积比值大小及数量的选择可以根据实际情况要求作出一些变动. 在没有特殊要求的情况下, 为了同时兼顾攻击成功率、最终扰动大小以及迭代次数, 本文在实验中选择了面积比值比为 1: 80、数量为 30 的矩形扰动块作为初始扰动.

3.6.2　交叉概率分析

交叉概率用来判定两个个体是否需要交叉, 其大小决定了进化过程的收敛速度以及收敛的优劣性. 本文利用 25 组车牌样本中的两个数字或字母测试了交叉概率对本实验的影响, 具体结果如表 7 所示.

表 7　交叉概率敏感性分析

Table 7　Cross-probability sensitivity analysis

交叉概率	迭代次数	原类标置信度	扰动大小 ($\tilde{L}_0$)
0.2	75	0.153	0.0048
0.4	53	0.138	0.0046
0.6	42	0.113	0.0048
0.8	34	0.126	0.0043
1	32	0.140	0.0045

由表 7 可知, 交叉概率为 0.2 时, 收敛时的迭代次数为 75 次; 交叉概率为 1 时, 收敛时的迭代次数下降到 32 次. 随着交叉概率的增大, 收敛时的迭代次数逐步下降. 但是交叉概率大于 0.8 时, 迭代次数的下降幅度大大减缓, 甚至几乎不再改变. 由此可知交叉概率对收敛速度具有较大的影响, 但大于 0.8 时影响细微. 由表 7 的后两列可知, 随着交叉概率的增大, 最优样本原类标置信度与扰动在一定范围内波动, 并没有体现出与交叉概率有较大的相关性, 所以最优对抗样本的鲁棒性与扰动大小对交叉概率不敏感. 为了节省时间成本以及保持样本的多样性, 本文选取交叉概率为 0.8.

3.6.3　迭代次数分析

在进化计算中, 迭代次数的多少直接体现该算法所需要的时间成本; 在黑盒攻击中, 迭代次数的多少直接决定了一次攻击需要访问的模型次数. 本文选取数据集中 10 张车牌样本, 分别对每张车牌上的 7 个字符进行攻击, 一共进行 70 次迭代攻击, 统计每 10 代最优种群 ($Rank_0$) 中的平均原类标置信度和平均扰动大小. 实验结果如图 9、图 10 所示.

由图 9 可知, 当迭代次数达到 35 代左右时, 最优种群 ($Rank_0$) 中的样本在平均原类标置信度上开

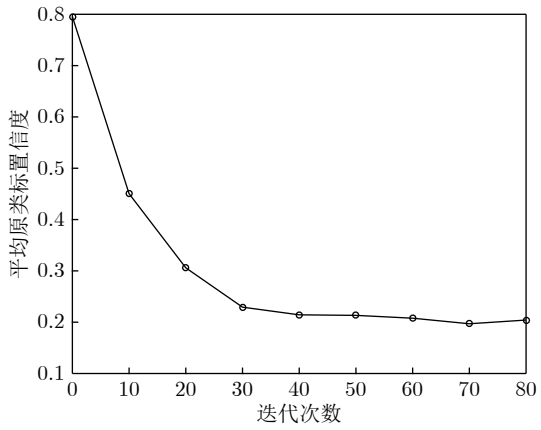


图 9 平均原类标置信度随迭代次数变化曲线图

Fig. 9 Curve of the original class standard confidence varying with the number of iterations

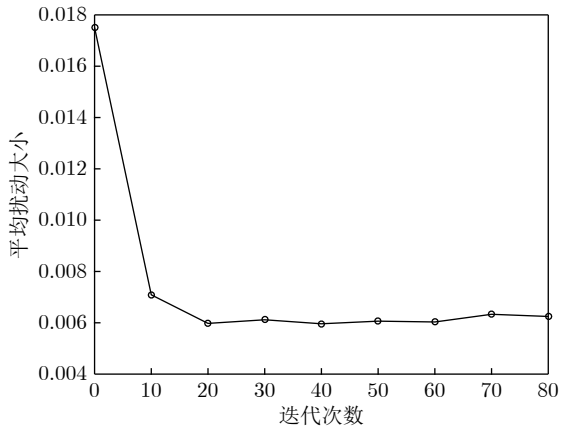


图 10 平均扰动大小随迭代次数变化曲线图

Fig. 10 Curve of the perturbation varying with the number of iterations

始收敛. 在前 30 代中, 随着迭代次数的增加, 种群中的平均原类标置信度下降速度变缓.

由图 10 可知, 当迭代次数达到 15 代左右时, 最优种群 ($Rank_0$) 中的样本在扰动大小上就已经开始收敛, 且前 10 代的下降速率很快, 表明对抗样本中扰动数的减少会在前 10 代产生质的变化.

结合图 9、图 10 可知, 最优种群 ($Rank_0$) 会在 35 代左右在两个目标函数上同时收敛. 在前 10 代中, 同时优化原类标置信度和扰动大小两个指标, 是进化过程中最重要的阶段; 在 10 代到 35 代中, 主要优化原类标置信度, 使得对抗样本的鲁棒性增强; 35 代以后的最优种群 ($Rank_0$) 达到收敛, 之后的迭代对种群优化影响不大. 所以本文算法不依赖大量的迭代次数, 大大减少了时间成本以及对目标模型的访问次数.

3.7 现实场景中商用车牌识别系统的攻击

最后, 为了更进一步证实本文的攻击方法在现实场景中的可实现性, 我们用泥土代替原先的扰动, 附在车牌的对应位置, 拍摄三种车牌识别场景下的车辆照片, 分别用百度 AI 开放平台中的车牌识别板块及具有商业性质的 OpenALPR 进行了检测, 验证本文算法的攻击具有迁移性和可实现性. 由于用泥土代替的扰动无法和打印出来的纸张扰动一样精准, 所以攻击效果较实验室环境有所下降. 本文设置了三种现实攻击场景, 分别为躲避公路上探头抓拍、躲避车牌尾号限行措施、冒充出入库车辆, 如表 8 ~ 表 10.

3.7.1 躲避公路上探头抓拍

本实验中正常车牌的识别结果为 “浙 A87065”, 生成扰动的目标模型为 HyperLPR. 本实验在 “8”、“0”、“6” 三个数字上添加扰动, 实现车辆躲避抓拍攻击. 由表 8 可知, 对于行驶过程中的车辆, 位置不同对识别结果具有较大的影响. 目标模型 HyperLPR 对中文字的识别效果较差, 但是对数字和字母的识别效果优于百度 AI 和 OpenALPR. 对于百度 AI 和 OpenALPR 而言, 三个添加扰动的数字均被错误识别.

3.7.2 躲避车牌尾号限行措施

本实验中正常车牌识别结果为 “苏 AN4D79”, 生成扰动的目标模型为 HyperLPR. 本实验在 “N”、“9” 上面添加扰动, 实现车辆的逃逸攻击以及躲避尾号限行攻击. 由表 9 可知, 对于添加了扰动的 “N”, 三个模型在不同位置均对 “N” 识别错误, 实现了现实生活中的车辆逃逸. 对于模型 HyperLPR, 尾号 “9” 均被识别为 “2”, 实现了现实生活中的躲避尾号限行. 对于其他两个模型, 大部分情况下, 尾号 “9” 也被错误识别, 证明针对于尾号的攻击也具有一定的迁移性.

3.7.3 冒充出入库车辆

本实验中正常车牌识别结果为 “浙 AP0P20”, 生成扰动的目标模型为 HyperLPR. 本实验在两个 “P” 上面添加不同的扰动, 实现车辆出入库顶替攻击. 出入库检测时, 拍摄角度较为倾斜, HyperLPR 模型无法正确检测出车牌的位置, 所以本文将其替换为学校出入库专用的商用车牌检测系统 (立方) 的检测结果. 由表 10 可知, 立方将一个 “P” 错误识别, 百度 AI 和 OpenALPR 将两个 “P” 都错误识别, 且三个商用软件都将第二个 “P” 错误识别为 “F”, 所以该扰动可作为迁移攻击的目标攻击.

表 8 躲避公路探头抓拍

Table 8 Avoiding road probe capture

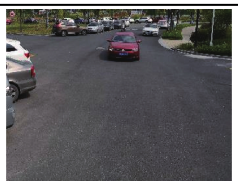
						
HyperLPR	云 AG7C35	HyperLPR	新 AG7C65	HyperLPR	浙 AG7C65	
百度AI	浙 AC7C35	百度AI	浙 AG7C35	百度AI	浙 AG7C35	
OpenALPR	浙 A67C65	OpenALPR	浙 A67C55	OpenALPR	浙 A07C35	

表 9 躲避车牌尾号限行

Table 9 Avoiding license plate tail number limit

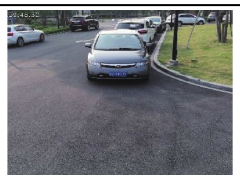
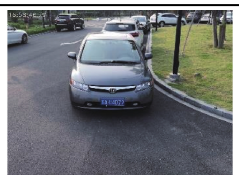
						
HyperLPR	苏 A14D72	HyperLPR	苏 AH4D72	HyperLPR	苏 AH4D72	
百度AI	苏 AH4D72	百度AI	苏 AH4D79	百度AI	苏 AH4D72	
OpenALPR	苏 AM4D78	OpenALPR	苏 AM4D79	OpenALPR	苏 AM4D72	

表 10 冒充出入库车辆

Table 10 Posing as a warehousing vehicle

						
立方	浙 AP0F20	立方	浙 AP0F20			
百度AI	浙 AF0F20	百度AI	浙 AT0F20			
OpenALPR	浙 A10F20	OpenALPR	浙 A10F20			

4 结束语

本文针对车牌识别系统提出了基于 NSGA-II 进化计算的黑盒物理攻击方法. 只需知道输出类标及对应置信度, 就能产生对环境变化因素具有较强鲁棒性的扰动, 而且本文算法将扰动控制为纯黑色块, 可用车子行驶过程中飞溅上来的淤泥块代替. 本文分别在实验室环境和真实环境中对生成的对抗样本进行检验, 验证了本文算法的物理可实现性以及对抗样本对真实环境因素的鲁棒性和迁移性.

除上述优点外, 本文算法也存在两个缺陷: 1) 生成的车牌扰动较大, 攻击可能会被外界因素所阻止 (如被交警拦下). 所以在之后的研究中, 设想用透明反光材料代替泥土, 大大降低人眼可见度. 2) 本文需要知道车牌模型输出的分类置信度, 在某些场合, 这个条件可能不被满足. 所以在之后的研究中, 尝试只用最终的分类类标进行攻击, 但是这可能会大大增加对模型的访问次数.

References

1 Goodfellow I J, Bengio Y, Courville A. *Deep Learning*. Cambridge: MIT Press, 2016. 24–45

2 Chen J Y, Zheng H B, Lin X, Wu Y Y, Su M M. A novel image segmentation method based on fast density clustering algorithm. *Engineering Applications of Artificial Intelligence*, 2018, **73**: 92–110

3 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada: MIT Press, 2014. 3104–3112

4 Dai Wei, Chai Tian-You. Data-driven optimal operational control of complex grinding processes. *Acta Automatica Sinica*, 2014, **40**(9): 2005–2014
(代伟, 柴天佑. 数据驱动的复杂磨矿过程运行优化控制方法. 自动化学报, 2014, **40**(9): 2005–2014)

5 Chen J Y, Zheng H B, Xiong H, Wu Y Y, Lin X, Ying S Y, et al. DGEPN-GCEN2V: A new framework for mining GGI and its application in biomarker detection. *Science China Information Sciences*, 2019, **62**(9): Article No. 199104

6 Yao Nai-Ming, Guo Qing-Pei, Qiao Feng-Chun, Chen Hui, Wang Hong-An. Robust facial expression recognition with generative adversarial networks. *Acta Automatica Sinica*, 2018, **44**(5): 865−877
(姚乃明, 郭清沛, 乔逢春, 陈辉, 王宏安. 基于生成式对抗网络的鲁棒人脸表情识别. 自动化学报, 2018, **44**(5): 865−877)

7 Yuan Wen-Hao, Sun Wen-Zhu, Xia Bin, Ou Shi-Feng. Improving speech enhancement in unseen noise using deep convolutional neural network. *Acta Automatica Sinica*, 2018, **44**(4): 751−759
(袁文浩, 孙文珠, 夏斌, 欧世峰. 利用深度卷积神经网络提高未知噪声下的语音增强性能. 自动化学报, 2018, **44**(4): 751−759)

8 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014). Banff, AB, Canada: ICLR, 2014.

9 Moosavi-Dezfooli S M, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 86−94

10 Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 2018, **6**: 14410−14430

11 Zeng X H, Liu C X, Wang Y S, Qiu W C, Xie L X, Tai Y W, et al. Adversarial attacks beyond the image space. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). California, USA: IEEE, 2019. 4302−4311

12 Deb K, Agarwal S, Pratap A, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 2002, **6**(2): 182−197

13 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, USA: ICLR, 2015.

14 Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). Toulon, France: ICLR, 2017.

15 Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2574−2582

16 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017. 39−57

17 Papernot N, McDaniel P, Jha S, Fredrikson M, Celik Z B, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy. Saarbrucken, Germany: IEEE, 2016. 372−387

18 Lyu C, Huang K Z, Liang H N. A unified gradient regularization family for adversarial examples. In: Proceedings of the 2015 IEEE International Conference on Data Mining. Atlantic City, USA: IEEE, 2015. 301−309

19 Su J W, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019, **23**(5): 828−841

20 Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, BC, Canada: ICLR, 2018.

21 Chen P Y, Zhang H, Sharma Y, Yi J F, Hsieh C J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM, 2017. 15−26

22 Tu C C, Ting P S, Chen P Y, Liu S J, Zhang H, Yi J F, et al. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019. 742−749

23 Chen J B, Jordan M I. Boundary attack++: Query-efficient decision-based adversarial attack. arXiv: 1904.02144, 2019.

24 Chen J Y, Su M M, Shen S J, Xiong H, Zheng H B. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. *Computers & Security*, 2019, **85**: 89−106

25 Bhagoji A N, He W, Li B. Exploring the space of black-box attacks on deep neural networks. arXiv: 1712.09491, 2017.

26 Chen S T, Cornelius C, Martin J, Chau D H. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Dublin, Ireland: Springer, 2019. 52−68

27 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137−1149

28 Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Tramer F, et al. Physical adversarial examples for object detectors. In: Proceedings of the 12th USENIX Workshop on Offensive Technologies. Baltimore, MD, USA: USENIX Association, 2018.

29 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 6517−6525

30 Thys S, Ranst W V, Goedemé T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE, 2019. 49−55

31 Sharif M, Bhagavatula S, Bauer L, Reiter M K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM, 2016. 1528−1540

32 Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1625−1634

33 Sitawarin C, Bhagoji A N, Mosenia A, Mittal P, Chiang M. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. arXiv: 1801.02780, 2018.

34 Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 284−293

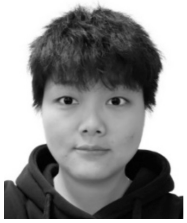
35 Li J C, Schmidt F, Kolter Z. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 3896−3904

36 Xu Z B, Yang W, Meng A J, Lu N X, Huang H, Ying C C, et al. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 261−277

陈晋音 浙江工业大学信息工程学院副教授. 分别于 2004 年, 2009 年获得浙江工业大学学士, 博士学位. 2005 年和 2006 年, 在日本足利工业大学学习进化计算. 主要研究方向为进化计算, 数据挖掘和深度学习算法. 本文通信作者.

E-mail: chenjinyin@zjut.edu.cn

(**CHEN Jin-Yin** Associate professor at College of Information Engineering, Zhejiang University of Technology. She received her bachelor and Ph. D. degrees from Zhejiang University of Technology in 2004 and 2009. She studied evolutionary computing in Ashikaga Institute of Technology, Japan in 2005 and 2006. Her research interest covers evolutionary computing, data mining, and deep learning algorithm. Corresponding author of this paper.)



沈诗婧 浙江工业大学信息工程学院硕士研究生. 主要研究方向为深度学习, 计算机视觉.

E-mail: 201407760128@zjut.edu.cn

(**SHEN Shi-Jing** Master student at College of Information Engineering, Zhejiang University of Technology. Her research interest covers deep learning and computer vision.)

苏蒙蒙 浙江工业大学信息工程学院硕士研究生. 2017 年获得浙江工业大学学士学位. 主要研究方向为智能计算, 人工免疫和工业安全.

E-mail: sumengmeng1994@163.com

(**SU Meng-Meng** Master student at College of Information Engineering, Zhejiang University of Technology. She received her bachelor degree from Zhejiang University of Technology in 2017. Her research interest covers intelligent computing, artificial immune, and industrial safety.)

郑海斌 浙江工业大学信息工程学院硕士研究生. 2017 年获得浙江工业大学学士学位. 主要研究方向为数据挖掘与应用, 生物信息学.

E-mail: haibinzheng320@gmail.com

(**ZHENG Hai-Bin** Master student at College of Information Engineering, Zhejiang University of Technology. He received his bachelor degree from Zhejiang University of Technology in 2017. His research interest covers data mining and applications, bioinformatics.)

熊 晖 浙江工业大学信息工程学院硕士研究生. 主要研究方向为图像处理, 人工智能.

E-mail: bearlight080329@gmail.com

(**XIONG Hui** Master student at College of Information Engineering, Zhejiang University of Technology. His research interest covers image processing and artificial intelligence.)