

针对车牌识别系统的双重对抗攻击

陈先意¹, 顾军¹, 颜凯¹, 江栋¹, 许林峰¹, 付章杰^{1,2}

(1. 南京信息工程大学数字取证教育部工程研究中心, 江苏 南京 210044;

2. 综合业务网理论及关键技术国家重点实验室(西安电子科技大学), 陕西 西安 710126)

摘要: 使用深度神经网络的人工智能系统面对对抗样本的攻击表现出极大的脆弱性。因此, 在基于深度神经网络的车牌识别系统(LPR, license plate recognition)场景下, 提出针对车牌识别系统的双重对抗攻击(DAA, double adversarial attack)方法。通过添加对抗补丁到车牌的图案位置, 使LPR的目标检测子系统不能检测到车牌类; 通过在车牌号上添加不规则的单连通区域扰动斑点以模拟自然形成的铁锈、污渍等, 使LPR的车牌号识别子系统产生误识别。针对车牌研究设计不同形状的对抗补丁、不同颜色的对抗斑点, 以此产生对抗车牌, 并迁移到物理世界中。实验结果表明, 设计出的对抗样本既不会被人眼所察觉, 也能够欺骗车牌识别系统, 如EasyPR, 在物理世界中的攻击成功率能够达到99%。关于LPR的对抗攻击及深度学习的脆弱性研究, 对提高车牌识别模型的鲁棒性具有积极促进意义。

关键词: 车牌识别; 对抗补丁; 对抗斑点; 对抗攻击

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023034

Double adversarial attack against license plate recognition system

CHEN Xianyi¹, GU Jun¹, YAN Kai¹, JIANG Dong¹, XU Linfeng¹, FU Zhangjie^{1,2}

1. Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China

2. The State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710126, China

Abstract: Recent studies have revealed that deep neural networks (DNN) used in artificial intelligence systems are highly vulnerable to adversarial sample-based attacks. To address this issue, a dual adversarial attack method was proposed for license plate recognition (LPR) systems in a DNN-based scenario. It was demonstrated that an

收稿日期: 2022-11-10; 修回日期: 2023-04-18

通信作者: 付章杰, www.fzj@126.com

基金项目: 国家重点研发计划(2021YFB2700900), 国家自然科学基金(62172232, 62172233), 江苏省杰出青年基金(BK20200039), 江苏省大气环境与装备技术协同创新中心基金

Foundation Items: The National Key R&D Program of China (2021YFB2700900), The National Natural Science Foundation of China (62172232, 62172233), The Jiangsu Basic Research Programs-Natural Science Foundation (BK20200039), The Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAET) Fund

引用格式: 陈先意, 顾军, 颜凯, 等. 针对车牌识别系统的双重对抗攻击[J]. 网络与信息安全学报, 2023, 9(3): 16-27.

Citation Format: CHEN X Y, GU J, YAN K, et al. Double adversarial attack against license plate recognition system[J]. Chinese Journal of Network and Information Security, 2023, 9(3): 16-27.

adversarial patch added to the pattern location of the license plate can render the target detection subsystem of the LPR system unable to detect the license plate class. Additionally, the natural rust and stains were simulated by adding irregular single-connected area random points to the license plate image, which results in the misrecognition of the license plate number. or the license plate research, different shapes of adversarial patches and different colors of adversarial patches are designed as a way to generate adversarial license plates and migrate them to the physical world. Experimental results show that the designed adversarial samples are undetectable by the human eye and can deceive the license plate recognition system, such as EasyPR. The success rate of the attack in the physical world can reach 99%. The study sheds light on the vulnerability of deep learning and the adversarial attack of LPR, and offers a positive contribution toward improving the robustness of license plate recognition models.

Keywords: license plate recognition, adversarial patch, adversarial spot, adversarial attack

0 引言

深度学习作为人工智能领域的一个重要分支,近年来被广泛应用于解决各种实际问题,其中,深度神经网络(DNN, deep neural network)作为最常用的机器学习方法,是一种高效的特征提取及建模工具,在许多计算机视觉任务中取得了超越人类极限水平的识别成果,如目标分类、人脸识别、智能驾驶、车牌识别(LPR, license plate recognition)等。此外,深度学习在日常生活中发挥着越来越重要的作用。

DNN 有效地加速了深度学习的发展,但是其自身存在的安全问题也日益凸显。多项研究证明,DNN 很容易受到对抗样本的攻击。2013年,Szegedy等^[1]在原始样本上添加了精心设计的、人眼难以发现的细微扰动,添加扰动后的图片(即对抗样本)使得包括卷积神经网络(CNN)在内的深度学习模型识别准确率明显降低,这意味着深度学习模型具有极大的脆弱性。

依据深度学习模型的内部结构是否公开,对抗攻击被分为白盒攻击和黑盒攻击^[2];依据攻击的具体实施场景,攻击又可被分为数字空间攻击和真实场景攻击^[3]。由于真实世界中的光照、天气等环境问题,从数字空间中产生的对抗扰动难以被打印出来,无法直接应用到真实世界中,因此真实场景下的对抗攻击更具挑战。

对抗样本的存在对于鲁棒性高或容错率高的应用系统是安全的,但是对于某些容错率要求严格的安全机要系统(如车牌识别系统),将会带来了巨大的安全隐患。车牌识别系统以车牌信息作

为唯一标识,通过使用训练好的车牌识别模型及车牌号识别模型,完成车牌的定位及车牌号码的识别。但是,车牌识别系统中存在的模型漏洞可能被不法分子利用从而躲避车牌监测^[4],致使整个识别系统形同虚设。因此,本文通过攻击车牌识别系统,挖掘深度学习模型的漏洞和脆弱点,对于提高车牌识别模型的鲁棒性具有重要意义。

本文针对车牌设计双重黑盒攻击,以证明车牌识别系统的潜在风险。目前针对车牌的真实场景攻击存在以下几个问题:① 对抗补丁面积大,易引人注目;② 补丁的颜色与形状等属性过于单一。考虑到攻击的伪装性和迷惑性,本文通过自定义车牌进行攻击。车牌中间是定制图案(如图1所示),四周是一些标语、日期标号等无关信息,车牌可进行定制、装饰,具有高度差异性。本文首先添加对抗补丁到与车牌信息无关的特定位置,然后在车牌字符上添加单连通区域扰动,模拟某种自然形成的斑点,最后将对抗补丁与对抗斑点集成,生成最终的对抗样本。实验结果表明,本文所设计的对抗样本不仅能够欺骗人眼视觉系统,具有一定的视觉伪装性,还能够有效地攻击车牌识别系统,以 EasyPR 为例,攻击成功率能够达到 99%。

综上所述,本文的主要创新点如下。

(1) 提出了一种单连通随机不规则形状区域生成(SCR, single connected random irregular shape region generation)算法,模拟车辆行驶过程中飞溅到车牌上的泥土或者其他自然形成的小型遮挡物(如铁锈),成功攻击车牌。

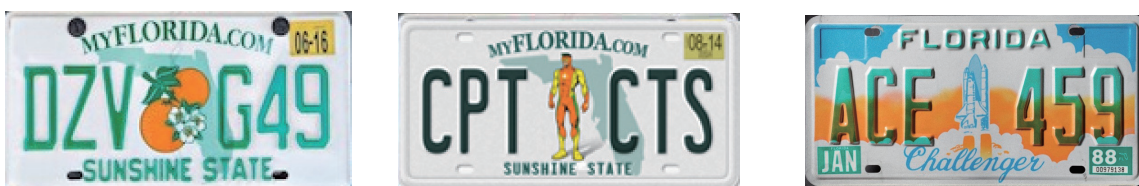


图 1 自定义车牌示例

Figure 1 Some examples of Custom license plate

(2) 提出双重对抗攻击 (DAA, double adversarial attack) 方法, 用于在目标检测和车牌号识别等不同推理环境下设计集成的对抗攻击。

(3) 针对当前最流行的目标检测框架 YOLOv5 及开源车牌号识别模型 OpenALPR、HyperLPR、EasyPR 及 PaddleOCR 进行攻击实验。结果表明, 在攻击 YOLOv5 等目标检测模型和 OpenALPR、HyperLPR 等车牌号识别模型时, 对抗样本攻击成功率极高。

(4) 提出在真实场景中车牌识别攻击方案, 具有较高的攻击成功率, 实现了对商用车牌识别系统的漏洞检测。

1 相关工作

1.1 针对分类任务的对抗攻击

2013 年, Biggio 等^[5]证明了对抗攻击的存在。之后, Szegedy 等^[1]成功地生成了针对分类模型的对抗样本, 他们通过在原始图像上稍微修改图像的像素值 (这样的修改并不会引起人眼的察觉, 却能够欺骗 DNN), 使 DNN 对图像进行错误的分类。在此之后, Goodfellow 等^[6]研究出一种快速符号梯度算法 (FGSM), 该算法能快速地生成图像扰动, 相比 Szegedy 等寻找最优受攻击图像的算法, 该算法目的是在更大的图像数据集中寻找能够对 DNN 进行攻击的单一图像。Moosavi-Dezfooli 等^[7]提出通过改变图像来产生攻击, 他们使用超平面来模拟不同输出类别与输入图像之间的边界, 并且攻击速度比前者更快。Carlini 等^[8]提出了另一种对抗攻击, 同样使用了优化方法, 与之前提到的攻击相比, 该算法在准确性和图像扰动前后相似性方面都有所提高。Brown 等^[9]提出了不改变像素值, 而是生成可以用数字方式放在图像上的补丁 (对抗补丁), 以欺骗分类器。Evtimov 等^[10]提出了一个

真实场景下的分类攻击, 他们将攻击对象定位于停车标志, 通过设计一张特殊的贴纸, 并将贴纸贴在停车标志上, 使其无法被自动驾驶识别。Athalye 等^[11]提出了一种优化三维模型纹理的方法, 考虑到物体的不同角度及光线变化, 以此设计优化函数, 最终用 3D 打印机打印出所得的物体, 这样的对抗样本具备鲁棒性。Moosavi-Dezfooli 等^[12]提出生成一个通用对抗样本, 该通用对抗样本被证明对不同的检测器都具有攻击能力。Tsai 等^[13]提出了一种针对 3D 点云分类网络的攻击方法, 该方法可以规避现有的专门针对对抗性三维数据的防御机制。

1.2 针对目标检测的真实场景攻击

目标检测本质上是一种多分类任务, 本文工作将从数字世界转移到真实物理世界。

Chen 等^[14]提出了针对目标检测的真实世界攻击方法, 该方法针对 Faster-RCNN 检测器^[15]中的停车标志实施攻击。与 Athalye 的工作类似, 他们使用了“期望变换” (EOT) 的概念 (在图像上做各种变换) 来建立对不同角度、姿势的鲁棒性。Evtimov 等同样以停车标志为攻击目标, 针对 YOLOv2^[16]设计白盒攻击, 在停车标志的整个红色区域内填入一个图案, 他们还在 Faster-RCNN 目标检测模型上进行了评估, 发现他们的攻击具备可迁移性。Thys 等^[17]提出在人身前拿着一个由对抗补丁打印而来的小纸板, 从而使目标检测器检测不到人的存在。Xu 等^[18]在 Thys 的基础上, 将对抗图案直接打印在 T 恤上, 并引入了薄板样条变换 (TPS) 以模拟 T 恤的褶皱, 生成的对抗 T 恤相比 Thys 中对抗 T 恤有着更高的攻击成功率。Kaziakhmedov 等^[19]通过普通的黑白打印机打印出该方法生成的各种不同的对抗补丁, 并将其印在医用口罩或将纸张直接贴在人脸上, 发现该方法能够在真实场景中破坏 MTCNN 检测器。

Huang 等^[20]研究了对野外目标检测器的真实场景对抗攻击, 提出学习一种对抗模式来有效地攻击属于同一对象类别的所有实例, 称为通用物理伪装攻击。Zhu 等^[21]提出了一种物理攻击方法, 使用热红外探测系统来攻击基于 YOLOv3 的行人检测器。Hu 等^[22]通过从 GAN 中采样最优图像, 可以在保持高攻击性能的同时生成视觉自然的对抗补丁。Hu 等^[23]提出了一种基于环形裁剪的可扩展生成攻击 (TC-EGA) 方法, 以制作具有重复结构的 AdvTexture, 将 AdvTexture 打印在 T 恤、裙子上, 实验表明, 这些衣服可以欺骗真实世界中的行人检测器。这个工作最令人瞩目的地方在于可以让对抗攻击不受角度的影响, 任意角度都可以攻击到, 但是衣服呈现出花花绿绿的颜色会有些奇怪。Wang 等^[24]将对抗图案印刷在汽车的车身上, 生成的对抗样本能够欺骗多种目标检测框架。Du 等^[25]将对抗补丁放置在汽车的车顶上, 使无人机或卫星等航空目标检测器出误。Wang 等^[26]在 Wang^[24]的基础上, 提出 FCA 涂装方案, 将车身的部分涂装拓展为全车涂装, 成功攻击 YOLOv3^[27]目标检测系统。Suryanto 等^[28]提出 DTA 框架, 利用了较新颖的 DTN, 增加了 3D 场景的参数控制, 实现涂装攻击。

1.3 YOLOv5

YOLO 算法最早由 Joseph Redmon 提出, 2020 年 5 月 30 日, YOLOv5 Ultralytics LLC 团队发布 YOLOv5, 通过 PyTorch 重写 Darknet 网络。YOLOv5 在 Tesla P100 快速检测上帧率可以达到 140, 而 YOLOv4^[29]帧率只有 50。同时, YOLOv5 的大小仅为 27 MB, 而使用 Darknet 架构的 YOLOv4 的大小为 244 MB, 且 YOLOv5 具有与 YOLOv4 相同的精度。YOLOv5 继承了 YOLOv4 的优点, 增加了 SPP-NET^[30], 提出了新的数据增强方法, 如 Mosaic training、SAT (self adversary training)、multi-channel 用 PANet^[31]替换 FPN 融合的特征。

1.4 车牌识别系统及车牌攻击

LPR 是现代智能交通系统的重要组成部分, 能够自动实时检测车辆通过和识别汽车牌照的智能交通管理系统, 它在交通管理、数字

安全监控、车辆识别和停车控制领域具有重要意义。

LPR 的基本流程如下: 首先对摄像头获取的车辆视频或者图像进行处理和分析; 接着利用模式识别技术或者数字图像处理技术得到车牌的相关信息。然而在实际生活场景中, 一些不确定因素, 比如拍摄车牌的光线明暗、天气恶劣与否、车辆运动速度等, 均会导致拍摄的车牌模糊。此外, 人为因素的遮挡或涂抹车牌等违法行为也会导致车牌识别的准确率较低。

为了提高 LPR 的识别速度与准确度, Parisi^[32]提出将 DNN 应用于 LPR。随着 DNN 的网络层数增多、结构变得更复杂, LPR 拥有了更高的识别准确率。基于 DNN 的车牌识别系统已经被广泛地使用, 如电子停车收费系统 (ETC)、智慧停车场系统、小区门禁系统、道路电子警察等。

常用的基于 DNN 的开源、商用车牌识别模型包括以下 4 种: HyperLPR、EasyPR、OpenALPR、基于百度飞桨的 OCR (Paddle OCR)。

目前, 针对 LPR 对抗攻击的工作并不多。Tsai 等^[13]在老式的佛罗里达州车牌上做了对抗补丁攻击。作者在这项工作中提出生成 3 种不同的对抗图案, 但每种图案的面积都过大, 在真实场景下给人的视觉效果并不理想。Qian 等^[33]提出斑点攻击, 以固定形状的斑点作为对抗补丁放置在车牌上, 以实现攻击的效果, 生成的补丁为固定形状, 较为不自然。Kwon 等^[34]提出在车牌上添加扰动的方法, 但是此方法仅适用于数字图像, 不能迁移到物理世界中。

2 DAA 方法

本文针对真实场景下的车牌识别系统, 基于自定义车牌, 提出了一种双重黑盒攻击方法, 旨在通过添加伪装补丁和微小斑点使得车牌识别系统识别错误。

车牌识别系统一般被认为是一个黑盒系统, 不能通过白盒方法攻击, 即根据其内部参数设计攻击。本文的攻击方法分为以下步骤: ① 使用 YOLOv5 训练出一个车牌目标检测模型, 替代车牌识别系统中的车牌检测模型, 替代模型的攻击

可以看作对车牌识别系统中车牌检测模型的迁移攻击；② 针对车牌上的标志性图案设计对抗补丁，用自定义的对抗补丁替换车牌上的图案；③ 使用单连通区域模拟任意可能情况下溅射到车牌上的不规则形状斑点、自然形成的铁锈等，生成最终的对抗斑点。步骤②中的对抗补丁目的是攻击车牌识别系统中的目标检测子系统，步骤③中的对抗斑点旨在攻击车牌号识别子系统。车牌识别系统双重对抗攻击方法的模型结构如图 2 所示。车牌识别系统双重对抗攻击方法包括两大模块：伪装对抗补丁攻击生成模块和不规则单连通区域斑点攻击生成模块。

2.1 伪装对抗补丁设计

本部分工作主要是生成抗 YOLOv5 车牌检测的对抗补丁。与 Brown 的工作不同（该工作的补丁放置位置会遮挡主图像），本文生成的补丁只覆盖了原来车牌上用于装饰的图案。本文设计了圆形、椭圆形和矩形 3 种补丁，实验验证各种形状下的补丁对检测器的攻击效果。

由于攻击不仅需要在电子图像上实现，还需要迁移到真实场景中，因此针对对抗补丁的损失优化包括 L_{nps} 、 L_{tv} 和 L_{obj} 这 3 个部分。

L_{nps} 指不可打印性损失^[17]，旨在使迭代出的补丁颜色能够正确地普通打印机打印出来。

$$L_{nps} = \sum_{p_{patch} \in P} \min_{c_{print} \in C} |p_{patch} - c_{print}| \quad (1)$$

其中， P_{patch} 指补丁 P 中的一个像素， C_{print} 是可被打印颜色集合 C 中的一种颜色。 L_{nps} 损失可以将图像上的颜色与可打印颜色集中的颜色尽可能相似，打印出来的补丁图案能够尽可能还原原图像。

L_{tv} 指平滑损失^[17]，描述图像的整体变化程度。这种损失能够指导优化器产生具有色彩平滑过渡的图像，防止了图像过度嘈杂。对补丁 P 计算 L_{tv} ，如式(2)所示。

$$L_{tv} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \quad (2)$$

计算相邻像素点之间的欧几里得距离，如果相邻像素相似，则 L_{tv} 的损失很低，如果相邻像素差异性过大，则 L_{tv} 损失很高。

L_{obj} 指检测对象损失，表示 YOLOv5 预测出图像中车牌的置信度得分，为此，训练的目标是尽量减少目标检测器输出的置信度。

联合不可打印性损失、平滑过渡损失及检测

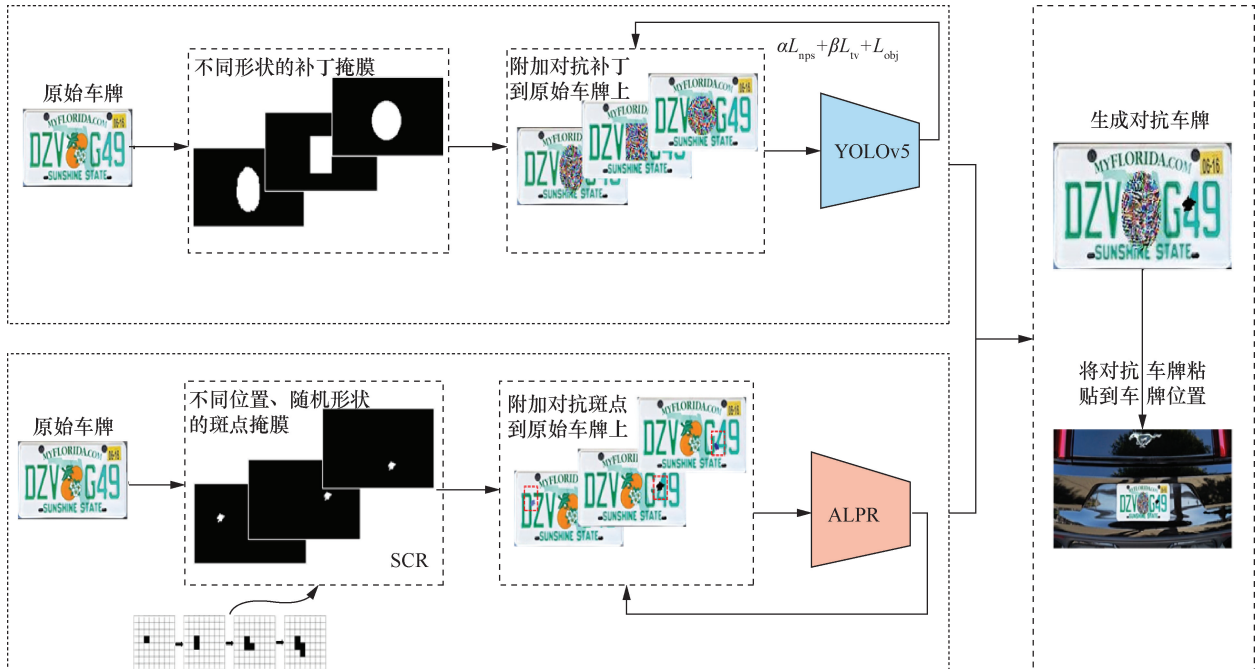


图 2 DAA 的模型结构
Figure 2 DAA model structure

对象损失来产生伪装性对抗补丁，具体如下。

1) 通过简单图像处理，从原车牌图像 I 中提取出目标补丁的位置，生成掩膜 $\text{patch}_{\text{mask}}$ ，如图3(a)所示。 $\text{patch}_{\text{mask}}$ 是一个矩阵，其中补丁区域的像素为1，其余区域像素为0，以此标记补丁的位置及形状。

2) 掩膜矩阵点乘原始图片矩阵，得到 I_{patch} ， I_{patch} 除了补丁区域外，像素皆为0。

3) 针对 I_{patch} ，使用梯度直接优化补丁，扰动受到 L_{nps} 和 L_{tv} 约束。

4) 将 I_{patch} 替换 I 中的对应部分，生成对抗样本 I_{adv} 。

5) 将 I_{adv} 送入已经训练好的 YOLOv5 目标检测模型中，获取预测值，根据 L_{obj} 的约束，重复2)~4)，直到攻击成功，即模型输出的车牌类置信度低或不再识别到车牌类别。

在这3个部分中，遵循如下的损失函数。

$$L = \alpha L_{\text{nps}} + \beta L_{\text{tv}} + L_{\text{obj}} \quad (3)$$

其中， α 控制 L_{nps} 的权重， β 控制 L_{tv} 的权重。本文中， α 和 β 设置为0.5。

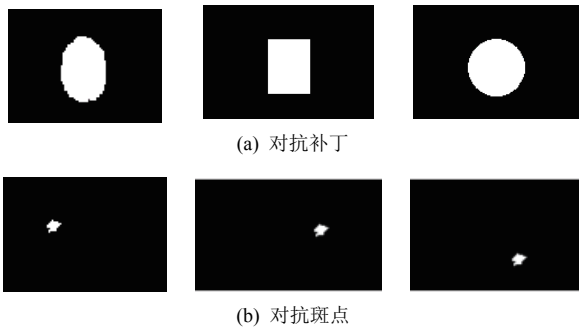


图3 对抗补丁和对抗斑点的不同掩膜
Figure 3 Different mask of adversarial patches and irregular spots

对抗补丁的训练算法可描述为算法1。

算法1 伪装对抗补丁生成算法

输入 原始图像 I ，原始图像中的补丁区域 I_{patch} ，对抗补丁 I'_{patch} ，基于 YOLOv5 的车牌识别模型 $R(\bullet)$ ，攻击的目标损失函数 $L(\bullet)$ ，超参数 α 、 β ，真实标签 l

输出 对抗样本 I_{adv}

过程

while $R(I_{\text{adv}}) = l$ do

 将 $I_{\text{patch}} + \text{sign}(\nabla I_{\text{patch}} J(I_{\text{patch}}, l))$ 赋值给

I'_{patch} ;

通过式(1)、式(2)计算 L_{nps} , L_{tv} 和 L_{obj} ;

通过式(3)优化 I_{adv} ;

end while

2.2 不规则单连通区域斑点攻击

本部分的工作独立于伪装补丁攻击，是针对图片上的“文字”进行的攻击。Kwon^[34]在车牌上添加的扰动可以使车牌识别系统失效，但是本文的工作并未迁移到物理世界中。Qian^[33]的攻击是在车牌的文字上添加固定形状的斑点，包括矩形、圆形及其他规则图案，但是这样的规则图案在现实场景中不自然。本文首次提出使用单连通区域生成算法来生成不规则斑点，这样的不规则斑点和汽车行驶时飞溅到车牌上的污渍相似。本文将对车牌号码的攻击建模为寻找最优的斑点颜色及位置。

算法2是生成不规则单连通区域算法的具体过程。

算法2 单连通区域生成算法

输入 宽为 W 、长为 L 的图片 I ，区域面积 N ;

输出 单连通区域 M_{required}

过程:

1. 在图片 I 中随机选取初始坐标 p
2. 将矩阵 M_{required} 使用0值初始化
3. 将点 $M_{\text{required}}[p]$ 加入边界数组 Boundary Array

4. while $N > 0$ do

 从 Boundary Array 中取随机点 $M_{\text{required}}[p]$;

 取 $M_{\text{required}}[p]$ 的邻接点，记作 $P_{\text{neighbours}}[1 \dots 4]$;

 取 $P_{\text{neighbours}}[1 \dots 4]$ 中值为0的任意一个点，将其值赋为1;

 对于 $P_{\text{neighbours}}$ 中的每个点，if 其4个邻接点值不全为1:

 将该点加入 Boundary Array

 对于 $P_{\text{neighbours}}$ 中的每个点，if 其4个邻接点值均为1且该点在 Boundary Array 中:

 将该点移出 Boundary Array.

 将 N 赋值为 $N-1$;

end while

5. 返回 M_{required}

对抗斑点的训练算法具体如下。

1) 指定单连通区域的像素值即面积 N ，在车牌的 MASK 区域，随机选择初始点 p ，生成单连通区域 spot，生成后的效果如图 3(b)所示。

2) 预设置 spot 的像素值，三通道像素值分别设置为 r 、 g 、 b 。

3) 将 spot 随机地贴到车牌图像 I 上，生成 I_{adv} 。

4) 将 I_{adv} 输入 LPR 系统中。

重复 2)~4)，直到预测值小于预设值 L ，停止迭代。

3 实验和分析

本文实验环境为一台装有 Intel Xeon Gold 6126 2.60 GHz CPU 和 32 GB 内存的 Centos7 服务器，由 4 颗 NVIDIA GeForce RTX3090 GPU 加速。用于评估的照片是由 iPhone11 Pro Max 的后置摄像头拍摄。

本文使用最先进的目标检测器 YOLOv5 评估了生成的物理对抗样本。YOLOv5 模型使用 Focus、BottleneckCSP 及 CSPNet^[35]作为骨干网络训练。对于伪装对抗补丁攻击，在互联网上搜集共 105 幅车牌图像，本文在 105 幅车牌图像的基础上做了旋转、模糊、明暗调整等数据增强技术，扩充了数据集的数量，最终获取共 420 幅图像作为数据集。基于上述数据集，训练出基于 YOLOv5 的车牌目标检测模型，只标记了一个类别即“车牌”类别，因为在这种攻击中，只需要考虑目标检测模型能否检测到车牌，其他类别与本实验毫无关系。对于不规则单连通区域斑点攻击，由于斑点的不规则和随机性，只限制斑点的面积与颜色。制造出的对抗车牌的尺寸是 15.24 cm×30.48 cm，将车牌打印到同样大小的纸张上，然后将其覆盖在车牌上。

本文评估了制造出的对抗车牌在各种条件下的效果，包括实验环境和真实场景。

3.1 识别模型

对车牌进行识别的开源模型包括以下 4 种：HyperLPR、EasyPR、OpenALPR、PaddleOCR。本文使用上述数据集对 3 种模型进行训练及测试，记录原始样本在不同识别模型下的识别准确率，如表 1 所示。

3.2 评价指标

为了评估所提方法的有效性，使用攻击成功率 (ASR, attack success rate) 作为评估指标。根据定义，ASR 是对抗攻击的成功百分比，即由于增加的扰动，攻击能够使模型预测错误的次数，在本文中指成功欺骗车牌识别模型的对抗样本数量与所有测试车牌数量之比。

$$ASR = \frac{\sum_{X'} \mathbb{I}((f_1(X') \neq y) \wedge (f_2(X') = y))}{\sum_X \mathbb{I}(f_2(X') = y)} \quad (4)$$

其中， X 表示原始车牌， X' 表示通过优化框架 DAA 生成的对抗车牌， y 表示车牌的真实标签， $f_1(\cdot)$ 表示基于深度 $\mathbb{I}(\cdot)$ 学习的识别模型， $f_2(\cdot)$ 表示人眼的判断结果，是指示函数。

表 1 对抗车牌在 4 种模型下的识别准确率
Table 1 Recognition accuracy of adversarial license plates under 4 models

模型名称	训练识别准确率	测试识别准确率
HyperLPR	95.7%	95.3%
EasyPR	89.4%	87.6%
OpenALPR	97.6%	96.8%
PaddleOCR	90.2%	87.9%

3.3 伪装对抗补丁攻击

对原始车牌添加不同形状的对抗补丁，包括圆形、椭圆形及矩形。如图 4 所示，第 1 行表示原始车牌，第 2 行是在原始的车牌图像上增添了圆形补丁，同样地，第 3 行和第 4 行分别对原始图像添加椭圆形补丁和矩形补丁。为了控制补丁形状这单一变量，约束 3 种形状的面积大小，以矩形为参照，圆形为矩形的外切圆，椭圆为矩形的内切椭圆，以此来比较出不同形状的补丁对车牌识别系统攻击成功率的影响程度。

表 2 表示不同形状的补丁随着攻击迭代次数增加后的攻击成功率。可以看出，随着攻击迭代次数的增加，3 种不同补丁形状的车牌攻击成功率也相应增加，这是由于随着迭代次数的增多，像素扰动增强。其中，圆形补丁的攻击成功率是 3 种形状中成功率最高的一种，这是因为圆形补丁的面积较其他形状大。需要说明的是，攻击迭



图 4 原始车牌（第 1 行）和添加了对抗补丁的对抗车牌（后 3 行）
Figure 4 Original license plate (first row) and plate with the patch added (last three rows)

代次数的增加势必会带来时间资源的消耗，时间消耗与迭代次数成正比。

表 2 不同形状的补丁随着攻击迭代次数增加后的攻击成功率

Table 2 The attack success rate of patches with different shapes varies with the number of attack iterations

补丁形状	攻击迭代次数				
	200	400	600	800	1 000
圆形补丁	14.25%	41.55%	72.24%	86.92%	91.85%
椭圆形补丁	12.37%	33.86%	61.23%	75.71%	84.13%
矩形补丁	7.32%	24.21%	45.69%	66.73%	75.11%

图 5 是不同迭代次数下，3 种补丁的攻击成功率。

图 6 为 Yang 等^[13]提出的 3 种对抗补丁生成方式，效果最好的为第二种，因为以这种方式生成的对抗补丁图案几乎充斥了整个车牌。在攻击基于 YOLOv5 的车牌识别模型时，在数字世界中达到 98.7% 的攻击成功率，在物理世界中达到 98.4% 的攻击成功率。

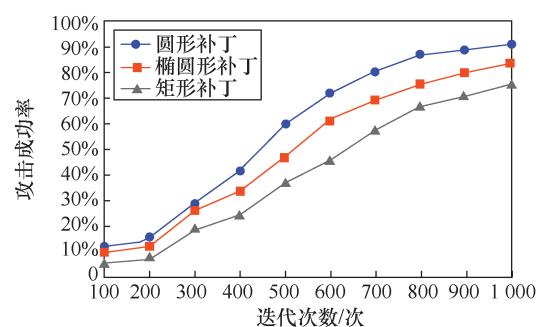


图 5 不同迭代次数下 3 种补丁的攻击成功率
Figure 5 Attack success rate with different iterations



图 6 Yang 等提出的 3 种对抗补丁
Figure 6 Three adversarial patterns proposed by Yang et al.

使用本文提出的 3 种不同形状的对抗补丁生成方法在经历 10 000 次迭代后，在数字世界中，攻击成功率的平均值达到 99.9%，物理世界中达到 98.4%，如表 3 所示。

表 3 攻击成功率的比较
Table 3 Comparison of attack success rates attack success

输入类别	攻击成功率	
	Yang 等	本文
数字世界	98.7%	99.9%
物理世界	98.2%	98.4%

3.4 不规则单连通区域斑点攻击

车牌识别系统中的号码识别部分主要是基于车牌号识别模型。通过生成不规则的单连通区域，模拟车牌上的污渍斑点，对车牌号识别模型造成攻击。

本文提出的斑点生成方法相较 Qian 等^[33]提出的斑点生成方法有很大的不同，Qian 等只能模拟出矩形、圆和椭圆等固定形状的斑点，但现实世界中，偶然溅射到车牌上的泥土污渍或铁锈不可能为固定形状。本文提出不规则单连通区域斑点生成算法，生成的斑点为随机形状，最大程度上模拟出物理场景中车牌的真实斑点。图 7 给出固定斑点和不规则单连通随机斑点。



(a) 固定形状的斑点 (b) 不规则单连通随机斑点
图 7 固定形状斑点和不规则单连通随机斑点
Figure 7 Fixed shape spots and irregular single-connected area random spot

图 8 表示在保持斑点颜色为黑色的情况下，不同面积的斑点对车牌号识别模型的影响， N 表示单连通区域的像素值大小。由图 8 可以得出结论，在对不规则单连通区域斑点的攻击情况下，斑点的面积越大攻击成功率越高。所生成的斑点对车牌图中的数字只有略微的遮挡，在其他的情况下对数字无遮挡。但是这样的不易让人察觉的斑点能够使车牌号识别模型产生接近 100% 的错误识别率。

3.5 数字场景中的双重攻击

将上述的伪装对抗补丁及不规则单连通区域

斑点攻击结合，实现对车牌的双重黑盒攻击，如图 9 所示。

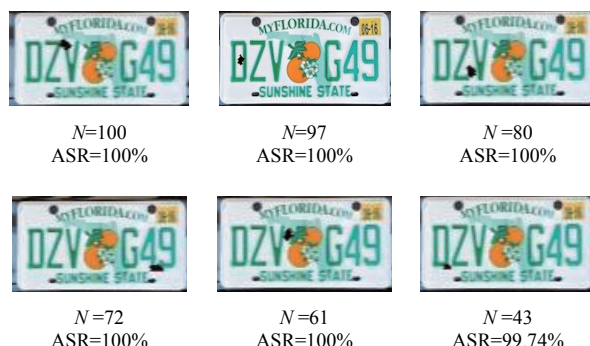


图 8 不同面积黑色斑点对车牌号识别模型的影响
Figure 8 Effects of black spots with different areas on license plate number recognition model



图 9 集成对抗补丁和对抗斑点后的对抗车牌
Figure 9 Adversarial license plate after integration of adversarial patch and spot

3.6 真实场景中的双重攻击

为了进一步证实本文的攻击方法在真实场景中的有效性，将上述结合了对抗补丁和对抗斑点的车牌对抗样本迁移到真实的物理场景下，即将此对抗样本图像按照真实车牌大小打印，并贴在汽车上。

本文拍摄了车牌识别场景下的车辆照片，原始车牌为 DZVG49，分别用 HyperLPR、EasyPR、PaddleOCR 和具有商业性质的 OpenALPR，验证本文算法的攻击具有迁移性和可实现性。

实验共拍摄了 100 组车牌图像，表 4 为 100 组真实场景车牌图像在 4 种模型下的攻击成功率。

表 4 真实场景中 4 种 LPR 模型的攻击成功率
Table 4 Attack success rate of four LPR models in physical world

模型	成功率
HyperLPR	98%
EasyPR	99%
PaddleOCR	94%
OpenALPR	98%

表 5 真实场景下攻击效果
Table 5 Attack performance of the proposed model in physical world

原始车牌	攻击场景 1		攻击场景 2		攻击场景 3	
						
DZVG49	HyperLPR	None	HyperLPR	DZVJ49	HyperLPR	DZV649
	EasyPR	None	EasyPR	None	EasyPR	DZV49
	PaddleOCR	DZG49	PaddleOCR	D2VG9	PaddleOCR	DZVG49
	OpenALPR	None	OpenALPR	None	OpenALPR	DZV49

实验场景及部分效果如表 5 所示,其中, None 表示此车牌识别系统未识别到车牌,原始车牌的车牌号为 DZVG49,其被识别为车牌的置信度为 0.89。通过实验可知,对抗补丁在一定程度上可以阻止 LPR 识别到车牌的存在;当 LPR 识别到车牌时,放置在车牌数字、字母上的不规则单连通斑点可以导致车牌号识别错误,如表 5 中 DZVG49 被误识别为 D2G49、DZVJ49 等。

实验中, PaddleOCR 的识别准确率较其他方法更高,可能是因为该模型本质上为 OCR 字符识别模型,较其他车牌识别模型有更高的字符识别能力。

4 结束语

本文针对车牌识别系统提出了双重攻击方法,具有重要的实用价值。未来的研究方向,应该是在深度神经网络安全方面进行更进一步的探索,以提高深度学习模型的鲁棒性,从而能够有效地抵御对抗攻击。期望未来的研究能够设计出更具有鲁棒性和可迁移性的对抗攻击方法,并能够在广泛的应用场景中保证深度学习模型的安全性和可靠性。同时,本文下一步工作将聚焦于真实物理世界中,多方位考虑其他物理环境因素可能对补丁和斑点造成的去噪效果,进一步提高攻击成功率,为深度学习模型的安全提供更加全面的保障。

参考文献:

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing

properties of neural networks[J]. arXiv Preprint arXiv:13126199, 2013.

[2] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020, 31(1): 67-81.

PAN W W, WANG X Y, SONG M L, et al. Survey on generating adversarial examples[J]. Journal of Software, 2020, 31(1): 67-81.

[3] 马钰锡, 张全新, 谭毓安, 等. 面向智能攻击的行为预测研究[J]. 软件学报, 2020, 32(5): 1526-1546.

MA Y Xi ZHANG Q X, TAN Y A, et al. Research on behavior-prediction for intelligent attacks[J]. Journal of Software, 2020, 32(5): 1526-1546.

[4] 陈晋音, 沈诗婧, 苏蒙蒙, 等. 车牌识别系统的黑盒对抗攻击[J]. 自动化学报, 2021, 47(1): 121-135.

CHEN J Y, SHEN S J, SU M M, et al. Black-box adversarial attack on license plate recognition system[J]. Acta Automatica Sinica, 2021, 47(1): 121-135.

[5] BIGGIO B, CORONA I, MAIORCA D, et al. Evasion attacks against machine learning at test time[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, 2013: 387-402.

[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint arXiv:14126572, 2014.

[7] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, America, 2016: 2574-2582.

[8] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017: 39-57.

[9] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. arXiv Preprint arXiv:171209665, 2017.

[10] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classifica-

- tion[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1625-1634.
- [11] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]//International Conference on Machine Learning. 2018: 284-293.
- [12] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.
- [13] TSAI T, YANG K, HO T Y, et al. Robust adversarial objects against deep learning models[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown, 2020: 954-962.
- [14] CHEN S T, CORNELIUS C, MARTIN J, et al. Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, 2018: 52-68.
- [15] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.
- [16] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.
- [17] THYS S, VAN R W, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.
- [18] XU K, ZHANG G, LIU S, et al. Adversarial t-shirt! evading person detectors in a physical world[C]//Computer Vision-ECCV 2020: 16th European Conference. 2020: 665-681.
- [19] KAZIAKHMEDOV E, KIREEV K, MELNIKOV G, et al. Real-world attack on MTCNN face detection system[C]//2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). 2019: 422-427.
- [20] HUANG L, GAO C, ZHOU Y, et al. Universal physical camouflage attacks on object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 720-729.
- [21] ZHU X, LI X, LI J, et al. Fooling thermal infrared pedestrian detectors in real world using small bulbs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 3616-3624.
- [22] HU Y C T, KUNG B H, TAN D S, et al. Naturalistic physical adversarial patch for object detectors[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7848-7857.
- [23] HU Z, HUANG S, ZHU X, et al. Adversarial texture for fooling person detectors in the physical world[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans Ernest N. 2022: 13307-13316.
- [24] WANG J, LIU A, YIN Z, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8565-8574.
- [25] DU A, CHEN B, CHIN T J, et al. Physical adversarial attacks on an aerial imagery object detector[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, New Orleans Ernest N. 2022: 1796-1806.
- [26] WANG D, JIANG T, SUN J, et al. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack[C]//Proceedings of the AAAI Conference on Artificial Intelligence., 2022: 2414-2422.
- [27] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. arXiv Preprint arXiv:180402767, 2018.
- [28] SURYANTO N, KIM Y, KANG H, et al. Dta: Physical camouflage attacks using differentiable transformation network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15305-15314.
- [29] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv Preprint arXiv:200410934, 2020.
- [30] PURKAIT P, ZHAO C, ZACH C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv Preprint arXiv:171203452, 2017.
- [31] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.
- [32] PARISI R, DI CLAUDIO E D, LUCARELLI G, et al. Car plate recognition by neural networks and image processing[C]//1998 IEEE International Symposium on Circuits and Systems (ISCAS). 1998: 195-198.
- [33] QIAN Y, MA D, WANG B, et al. Spot evasion attacks: Adversarial examples for license plate recognition systems with convolutional neural networks[J]. Computers & Security, 2020, 95: 101826.
- [34] KWON H, BAEK J W. Adv-plate attack: adversarially perturbed plate for license plate recognition system[J]. Journal of Sensors, 2021, 2021: 1-10.
- [35] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 390-391.

[作者简介]



陈先意（1983- ），男，湖北恩施人，南京信息工程大学副教授，主要研究方向为区块链安全、大数据安全及人工智能安全。



江栋（1998- ），男，四川成都人，南京信息工程大学硕士生，主要研究方向为人工智能安全、对抗样本。



顾军（1996- ），男，江苏盐城人，南京信息工程大学硕士生，主要研究方向为对抗样本、人工智能安全。



许林峰（1998- ），男，江苏南通人，南京信息工程大学硕士生，主要研究方向为信息安全、对抗样本。



颜凯（1998- ），男，江苏淮安人，南京信息工程大学硕士生，主要研究方向为对抗攻击。



付章杰（1983- ），男，河南南阳人，南京信息工程大学教授、博士生导师，主要研究方向为区块链安全、数字取证、人工智能安全。