一种基于二维码对抗样本的物理补丁攻击

钱亚冠[1], 刘新伟[1], 顾钊铨[2], 王 滨[3], 潘 俊[1], 张锡敏[1]

[1]浙江科技学院大数据学院, 杭州 中国 310023
[2]广州大学网络空间先进技术研究院, 广州 中国 510006
[3]杭州海康威视网络与信息安全实验室, 杭州 中国 310051

摘要 深度学习技术在图像识别领域已经得到广泛应用, 识别准确率超过人类平均水平。然而最近的研究表明, 深度神经网络的性能会因对抗样本的存在而大幅降低。攻击者通过在待识别的图像中添加精心设计的微小扰动, 误导分类器做出错误预测。另一个方面, 在数字空间生成的扰动也能够转移到物理空间并用于攻击。为此, 本文提出了一种基于二维码对抗样本的物理补丁攻击方法。将生成的二维码贴在道路交通标志表面的指定位置, 使得分类器输出错误的分类。实验结果表明了本文方法的有效性, 同时, 将数字空间生成的对抗样本用于物理空间中的交通标志攻击, 仍可以保持较高的成功率。

关键词 深度学习; 对抗样本; 二维码; 补丁攻击
中图法分类号 TP309 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.11.07

QR Code Based Patch Attacks in Physical World

QIAN Yaguan[1], LIU Xinwei[1], GU Zhaoquan[2], WANG Bin[3], PAN Jun[1], ZHANG Ximin[1]

[1] School of Sugon Big Date Science, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2] Cyberspace Institute of Advanced Technology (CIAT), Guangzhou University, Guangzhou 510006, China
[3] Network and Information Security Laboratory of Hangzhou Hikvision Digital Technology Co, Ltd. Hangzhou 310051, China

Abstract Deep learning technology has been widely used in the field of image recognition, and the recognition accuracy is higher than the average level of human beings. However, recent studies have shown that the performance of deep neural network will be greatly reduced due to the presence of adversarial examples. The attacker misleads the classifier to make false prediction by adding a small disturbance to the image to be recognized. On the other hand, the disturbance generated in the digital space can also be transferred to the physical space and used for attack. For this reason, this paper proposes a physical patch attack method based on two-dimensional code antagonism samples, which pastes the generated QR code on the designated position of the road traffic sign surface, making the classifier output the wrong classification. The experimental results show the effectiveness of this method. At the same time, using the counter examples generated in digital space to attack traffic signs in physical space can still maintain a high success rate.

Key words deep learning; adversarial examples; QR code; patch attack

1 引言

近年来, 深度神经网络(Deep Neural Networks, DNN)被广泛应用于人脸识别[1]、车牌识别[2]、自然语言处理[3]等众多领域, 并取得了巨大成功。然而, 随着技术的不断发展, 深度神经网络模型也暴露出对抗样本[4]攻击等诸多安全问题。研究表明, 如果攻击者在样本数据中引入不易察觉的、经过特别设计的扰动, 可能导致深度学习模型做出错误的分类判断。这表明深度学习模型的安全性需要引起重视, 许多因素会影响模型做出正确判断, 进而导致不必要的损失。研究表明, 对抗样本能够对无人驾驶车辆的识别系统进行攻击, 例如, 科恩实验室的研究人员通过引入地面干扰信息, 成功误导特斯拉无人驾驶汽车驶入反向车道[5]。需要指出的是, 在数字空间生成的攻击样本, 已经适用于物理空间中的攻击场景[6]。本文以道路交通标志攻击的对抗样本为研究对象, 旨在证明这种攻击的可行性, 为防御该类攻击作好前期准备工作。

对抗样本是深度神经网络安全研究的一个重要

通讯作者: 钱亚冠, 博士, 副教授, Email: qianyaguan@zust.edu.cn。
本课题得到科技部重点研发项目(No.2018YFB2100400); 国家自然科学基金资助项目(No.61902082); 浙江省自然科学基金资助项目(No.LY17F020011); 浙江省公益技术应用研究项目(No.LGF20F020007, No.LGG19F030001)资助。
收稿日期: 2019-12-29; 修改日期: 2020-01-05; 定稿日期: 2020-09-22

组成部分。针对图像识别任务, 生成对抗样本的攻击方法主要有两类。一类是针对图像全域像素的对抗攻击, 例如 Goodfellow 等人[7] 提出了一种快速梯度符号法(fast gradient sign method, FGSM)。通过创建沿梯度相反方向的扰动来误导分类器。FGSM 速度较快, 但每次攻击只涉及单次梯度更新, 容易陷入局部最优值。因此, 一些研究人员在 FGSM 的基础上作出了各种改进, 最近 Cihang Xie 等人[8]提出了 M-DI^2-FGSM 方法, 构造了更加精准的扰动, 从而提高攻击成功率并缓解了陷入局部最优的问题。另一类是对图像局部像素的对抗攻击, 例如 Papernot 等人[9]提出了一种针对有目标攻击的迭代方法(Jacobian-based Saliency Map, JSMA): 用分类器分类得到的 logits 值来计算样本像素点的梯度, 依次迭代得到显著性图(Saliency map), 通过显著性图来生成对抗样本。与 FGSM 及其延伸的方法相比, JSMA 攻击中所需要扰动的像素点较少, 但显著性图搜索计算开销巨大, 不适用于三通道的彩色图片优化求解[10]。Su 和 Vargas 等人[11]提出了单像素攻击方法, 只需修改单个像素的值, 就能成功欺骗分类器, 但这种方法只适用分辨率较低的图片数据, 如 MNIST 数据集和 CIFAR 数据集, 对于高分辨率的彩色图片, 其攻击性能会大幅降低。

上述两类方法都可通过优化算法对图像像素施加扰动, 在数字空间中生成对抗样本, 添加扰动后的图像通常也不易被人们发觉。然而, 在将数字空间生成的对抗样本转移到物理空间时, 会因光照、角度、摄像机拍摄距离的不一致等因素导致攻击失败。为此, 本文提出一种使用二维码贴纸来攻击道路交通标志图像的方法。首先通过分类器计算得到道路交通标志图像的特征图[12], 确定二维码补丁贴纸的攻击位置, 然后对二维码贴纸内的数据区域进行优化计算, 最终生成对抗样本。我们分别在数字空间和物理空间中展开实验, 实验结果表明, 在数字空间中攻击成功的二维码补丁, 在物理空间中依旧可以保持较高的攻击成功率。本文的创新点和贡献如下:

(1)现有的对抗样本攻击往往以图像全域像素为攻击对象, 而本文以二维码补丁作为贴纸进行攻击, 仅需修改二维码补丁内的像素区域, 需要计算的面积更小, 速度更快。

(2)以真实的道路交通标志作为攻击对象, 分别在数字空间和物理空间中进行测试, 实验结果表明, 对抗样本在物理空间中依旧保持较高的攻击成功率。

(3)本文以现实生活中常见的二维码作为攻击补丁, 具有很高的迷惑性, 不易被察觉, 而且生成二维码对抗样本的成本低, 实施攻击更符合现实场景。

2 背景知识和相关工作

2.1 卷积神经网络与对抗样本

深度神经网络(DNNs)分类器可以表示为映射函数: $F(x,\theta):\mathbb{R}^d\to\mathbb{R}^L$,其中 $x\in\mathbb{R}^d$ 是输入变量, θ 表示所有的模型参数, L 则代表输出类的个数, 最后一层神经网络层一般采用 Softmax 层, 可定义为函数:

$$S(z)_i=\exp(z_i)\Big/\sum_{i=1}^{L}\exp(z_i),i\in[L],[L]=\{1,\cdots,L\} \quad (1)$$

设 Z 为最后一个隐藏层的输出向量, 定义一个映射函数 $x\to Z$ 来表示提取数据, DNNs 则可以表示为: $F(x)=S\times(W_S Z+b_S)$, 其中 W_S 和 b_S 分别 Softmax 层的权重和偏置向量, $W_S Z+b_S$ 称为 logits。给定一个输入 x 得到预测标签为 $y=\arg\max_{i\in[L]}F(x)_i$, 概率值 $F(x)_y$ 作为置信度。DNNs 的训练目标是使交叉熵(Cross-entropyloss)的损失最小化, 其可定义为:

$$l_{ce}(x,y)=-O_y^T\log F(x)=-\log F(x)_y \quad (2)$$

每一组输入的 (x,y) 中, O_y 是对 y 标签进行 One-hot 编码, $\log F(\cdot)$ 则表示对每个元素取对数, 交叉熵损失最小化的目的是为了获得最佳的训练参数。

研究表明, DNNs 很容易受到对抗样本的攻击。给定一个分类器 $F(x)$, 其中 $x\in X,y\in Y$, 输入一个样本 x 得到一个对应预测标签 y, 真实标签为 y_{true}。Szegedy 等人首次在文献[4]中提出了关于对抗样本(Adversarial examples)的概念: 对正常样本 x 添加受范数约束的不超过 ε 的扰动得到新的样本 x', 使得 DNNs 以高置信度给出一个错误的预测标签 y^*。寻找对抗样本的过程可以视为一个优化求解问题:

$$\max_{x'}F(y=y^*\mid x'),\quad s.t.\quad \|x'-x\|_p\le\varepsilon \quad (3)$$

其中 p 可以取值为 $0,2,\infty$, 通过限制 L_0、L_2 和 L_∞ 范数使得添加的扰动无法被察觉。

2.2 威胁模型

在分类问题中, 样本 x 由 DNNs 正确分类得到真实标签 y_{true}。当对抗样本 x' 经过 DNNs 分类后得到 $y'\neq y_{true}$ 时代表攻击成功, y^* 为指定目标类别, 且 $y^*\neq y_{true}$。当 $y'=y^*$ 时称为有目标攻击(targeted attacks), 反之 $y'\neq y_{true}\neq y^*$ 时则为无目标攻击

(untargeted attacks)。有目标攻击往往比无目标攻击更加有难度。文献[13]中定义了三种威胁模型存在:

(1)零知识(An Zero-Knowledge Adversary): 对于模型参数等没有任何的了解, 对抗样本 x' 在本身并不鲁棒的模型 F 中产生, 并且对于 DNNs 信息没有任何的了解; (2)完全知识(A Perfect-Knowledge Adversary): 攻击者知道 F 的模型结构、训练数据、参数等详细信息的前提下, 生成对抗样本 x', 即白箱攻击; (3)有限知识(A Limited-Knowledge Adversary): 攻击者无法获得已训练完成的神经网络模型 F 的内部信息, 但能获得输入对抗样本 x' 后输出的标签和置信度等信息。即黑箱攻击。

对于攻击者而言, 虽然模型在训练完成之后进行了封装, 但是依旧可以通过大量样本返回的信息嗅探模型的内部结构, 在本地搭建一个代理模型进行攻击。

2.3 物理攻击方法

引言中介绍了数字空间下的几种攻击方法, 但是数字空间中产生的对抗样本在物理空间中因为相机拍摄角度、光的反射、图片不清晰等原因可能失效, 所以, 如何将数字空间中的扰动应用到物理空间中是一个难处理的问题。

Kevin Eykholt 等人[14]从优化方法推导出一种新的方法, 该方法为单个图像 x 生成扰动, 在不考虑其他物理条件下, 输入图像 x 需要添加的扰动为 δ, 这样的扰动实例 $x'=x+\delta$ 由目标分类器 f_θ 分类:

$$\arg\min_{\delta}\lambda\|\delta\|_p+J(f_\theta(x+\delta),y^*) \tag{4}$$

其中, y^* 是目标类, J 是损失函数用于衡量模型预测和目标标签 y^* 的差异, λ 是一个超参数用于控制失真的正则化, 距离函数 $H=\|\delta\|_p$, 表示 δ 的 l_p 范数。这种方法中生成的扰动补丁占输入样本面积的比例非常大, 而且通常需要多块补丁联合才能攻击成功。

Sitawarin 等人[15]受到 Adam 优化器[16]来解决非凸优化问题和 C&W 攻击方法[17]的启发, 对于任意输入样本 x, 鲁棒的对抗性样本可以表示为以下给出问题的最优解:

$$\begin{aligned}&\min_{\delta\in\mathbb{R}^n}\quad c\cdot\frac{1}{B}\sum_{i=1}^{B}[F(\tau_i(x+M\cdot\delta))]+\max(\|\delta\|_p,L)\\&s.t.\ F=\max(\max\{Z_{pre}(x)_i:i\neq y\}-Z_{pre}(x),-K)\end{aligned} \tag{5}$$

其中, K 用来调节目标分类的置信度, τ_i: $\mathbb{R}^n\to\mathbb{R}^n$ 是变换函数在图像空间中的映射, M 是与输入图像 x 具有相同宽度和高度的 0 或 255 像素值矩阵的掩膜, 而 $M\cdot\delta$ 的目的是为了将可行区域约束到 Sign area。常数 c 平衡真实目标函数和惩罚项, 常数 L 是为了防止扰动过小而导致无法被相机捕捉, 其中规定扰动的范数至少为 L。实验中设置 L 值为 1, 尽管 L 值越大, 攻击成功率越高, 但过大的扰动也容易被人眼察觉, 所以实验中设置为 1 比较合适。该方法需要在整张图片样本上迭代搜索, 导致得到对抗样本需要花费大量的时间。

3 基于二维码的补丁攻击

本文研究了 2.3 中物理空间下的攻击方法后发现, 这些方法虽然能够产生有效的对抗样本, 但是在制造对抗样本的过程中往往存在以下几种缺陷: (1)以整个路标图像作为攻击对象, 导致生成对抗样本需要大量的时间; (2)扰动补丁占整个路标区域的面积较大, 文献[14]中使用的攻击补丁约占整个路标面积的 4.8%, 且需要多块不同颜色的补丁进行搭配才能产生一定的攻击效果;(3)计算生成的扰动补丁颜色明显, 没有具体的图案特征, 容易引起人们的注意, 暴露攻击意图。

为了克服以上的几种缺陷, 我们参考特征图[12]选择攻击位置, 利用二维码补丁隐蔽性强的特点进行攻击。本文的方法具有以下优点: (1)攻击的二维码补丁面积较小, 攻击成功的最小补丁约占道路交通标志图像面积的 0.8%, 平均面积 0.95%, 相比文献[14]中攻击补丁面积减少了 80.21%;(2)该方法只对道路交通标志中的二维码补丁进行优化计算, 可以更快生成对抗样本;(3)二维码为生活中非常常见的图案, 用其制作对抗样本不易引起人们的警觉, 且制作简单, 易于实现。

3.1 二维码生成原理

二维码[18]是按一定规律在二维平面上分布的黑白或彩色相间的图形, 用于表示某种符号信息。二维码的生成规则借鉴了计算机中二进制编码, 用多个二进制对应的矩形几何形状来存储表达文字数值信息。常用的二维码码制有: Data Matrix, Maxi C-ode, Aztec, QR Code, Vericode, PDF417, Ultracode Code 49, Code 16K 等。由于 QR 码有 3 处定位图案, 不受背景样式的影响, 可以从任一方向均可快速读取, 因此我们选择 QR 码制作为本文生成二维码补丁的编码规则。QR 码设有 1 到 40 的不同版本, 每一个版本比前一个版本增加 4 个码元, 计算公式为 $(n-1)\times4+21$, 其中 n 为版本号。每个版本都具备固有的码元结构。码元是指构成 QR 码的方形黑白或彩

色点, “码元结构”是指二维码中的码元数, 从版本 1(21 码元×21 码元)开始, 在纵向和横向各自以 4 码元为单位递增, 一直到版本 40(177 码元×177 码元)。每个码元存储一个二进制的 0 或者 1, 其中黑白的二维码中 0 代表黑色的码元, 1 则表示白色的码元。

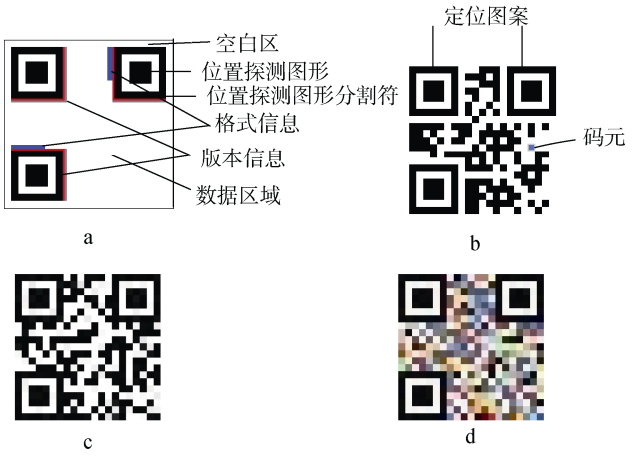

图 1　二维码示例

Figure 1　QR code example

本文中定义了一种三通道的黑白或彩色伪二维码 M, 与普通二维码不同的地方在于剔除了格式和版本信息, 但均包含有定位图案。具体如图 1 所示, a 图是普通黑白二维码的结构示意图, 空白区是为了二维码便于标识, 位置探测图形(也叫定位符)是由三个黑白相间的大正方形嵌套组成, 分别位于二维码左上角、右上角、左下角, 目的是为了确定二维码内数据区域的位置。位置探测分隔符是为了将位置探测图形和数据区域进行分割, 格式信息用来记录使用的掩码和纠错等级。版本信息仅在版本 7 以上存在, 记录具体的版本信息。数据区域记录的是由码元组成的数据, 在本文中伪二维码仅保留数据区域和定位符。b 图是由 21 码元×21 码元生成的普通二维码图案。c 图则是由 21×21 码元生成的不包含格式版本信息的三通道黑白伪二维码图案。d 图是由 21×21 码元生成的不包含格式版本信息的三通道彩色伪二维码图案。

3.2　特征图

文献[12]中对特征图给出了定义: 一张大小为 $m\times n$ 的图片样本 x, 对应的正确分类是 y_{true}, 分类器对其分类到 y_{true} 的概率值是 S, 通过特征图来衡量 x 中的某个像素点对于该分类器分类到 y_{true} 的概率值 S 的影响。计算特征图需要计算与输入样本像素对应的正确分类中标准化分数的梯度, 它用来衡量图像中的某个像素点发生轻微变化时, 正确分类的分数变化幅度。因此, 可以借助该方法来确定伪二维码补丁添加在道路交通标志图像上的具体坐标。在分类器对道路交通标志图像进行分类的同时计算得到特征图, 在这些像素点组成的特征图的像素点中选取对分类器分类影响最大的坐标点作为伪二维码补丁添加到路标的具体位置, 效果分别如图 2, 图 5 所示, 其中图 2 是部分路标图像经过 DNNs 正确分类之后得到的特征图, 图 5 是在获得特征贴图之后确定伪二维码添加坐标, 得到初始对抗样本。

图 2　部分路标图像和对应的特征图示例

Figure 2　Example of part of road sign image and corresponding feature map

3.3　攻击方法

本文将路标样本 x 输入 DNNs 得到预测目标类置信度值记为 $F_y(x)$, 其真实标签为 y_{true}。求解最小扰动 ε 的前提下生成令 DNNs 产生错误分类的对抗样本 $x'=x+\varepsilon$。具体的过程可以转化为下式:

$$\min\ d(x,x+\varepsilon),\qquad s.t.\quad F_t(x+\varepsilon)=y^* \tag{6}$$

其中, d 是一个可选择的距离函数, F_t 是目标分类器, y^* 是目标类。这样的方法在数字空间中虽然是可行的, 但由于受到物理空间中光照、角度、拍摄距离等不确定因素的影响, 数字空间生成的对抗样本在物理空间中难以保证其有效性。为了更好的适应物理世界的复杂变换, 使数字空间中产生的对抗样本在物理空间仍有效, 我们结合式(4), 提出如下解决方案:

$$\arg\min_{\varepsilon}\ l(F(x+t(M\cdot\varepsilon)),y^*),\quad s.t.\quad \max(\|\varepsilon\|_p,d) \tag{7}$$

其中, l 是损失函数用于衡量模型预测与目标标签 y^* 的差距。$t_i:\mathbb{R}^n\to\mathbb{R}^n$ 是变换函数在空间中的映射(例如: 当使用平移、翻转、旋转等几何变换对样本图像进行数据增强时, 扰动也随之进行变换)。d 是一个常数, 使得 $\|\varepsilon\|_p$ 至少为 d 用于防止攻击补丁过小的扰动在物理空间中因打印和视频捕捉后较为模糊等问题而失效。在户外测试中, 我们设定最远测试距离为 d =15 m, 实验中可以得到清晰的二维码, 我们以此二维码的大小来设计伪二维码。为了有效的

解决上述约束优化问题, 我们用拉格朗日乘子的形式将上式进行重新表达:

$$\arg\min_{\varepsilon} \; l(F(x+t(M\cdot\varepsilon)),y^*)+\lambda(d-\|\varepsilon\|_p) \tag{8}$$

λ 是正则化系数, 用于控制扰动失真和损失函数之间的比例。$(M\cdot\varepsilon)$ 表示伪二维码, 在加入伪二维码补丁的路标图像中对其进行如下优化, 得到伪二维码补丁攻击算法:

$$x'_{N+1}=Clip_{x,\varepsilon}\left\{x'_N+\delta\cdot\mathrm{sign}\left\{\nabla_t\left[l(F(x+t(M\cdot\varepsilon)),y^*)\right]\right\}\right\} \tag{9}$$

δ 表示权重, N 代表迭代次数, $Cilp$ 表示将溢出的数值用边界值代替来防止在更新中因迭代次数的增加, 部分像素值可能会溢出(超出正常像素值的 0 到 255 的范围), 此时将小于 0 的值用 0 替换, 大于 255 的值用 255 替换, 从而保证可以生成有效的图像。

如图 3 所示, 以向左急转弯道路交通标志为例来说明伪二维码补丁攻击流程。(1)将图片样本 x 经过 DNNs 分类器分类, 计算分类到正确图片类别的特征图, 并返回样本 x 对 DNNs 分类器最大影响点的坐标 $E(a,b)$; (2)程序随机生成一个伪二维码 M, 并将 M 的中心坐标点贴在点 $E(a,b)$ 得到初始对抗样本 x'; (3)将 x' 由 DNNs 分类器分类, 若分类出错则将 x' 作为对抗样本 x_{adv} 保存, 表示攻击成功。否则将伪二维码由伪二维码补丁攻击算法进行优化, 生成新的伪二维码后重复以上过程, 直至攻击成功得到对抗样本 x_{adv}, 或达到最大迭代次数后仍无法得到对抗样本 x_{adv}, 表示攻击失败。具体过程由算法 1 实现。为了更贴近现实环境以及提高运算效率, 实验中将伪二维码中的每一个码元缩小至单个像素大小, 同时也将交通标志图片样本等比例缩小。在像素级别下, 我们可以直接使用式(9)对伪二维码之内的数据区域进行优化求解, 这样大大减少了计算量, 并在短时间内获得了对抗样本 x_{adv}。

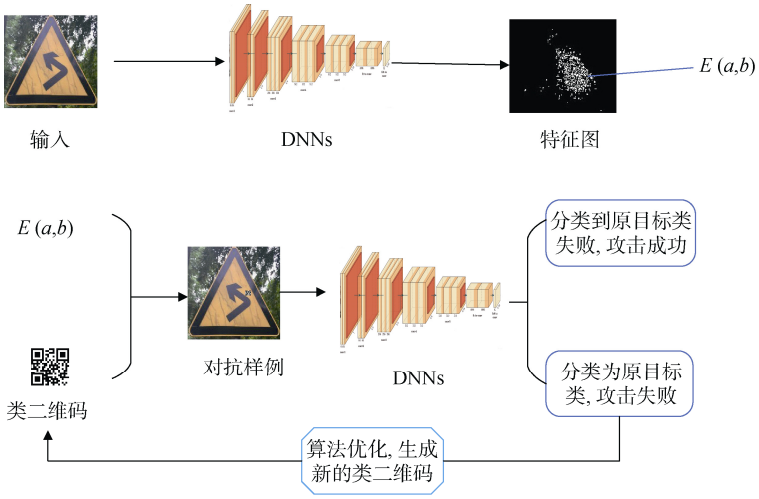

图 3 伪二维码攻击示意图

Figure 3 Pseudo QR code attack

算法 1. 生成对抗样本 x_{adv}

输入: 路标图片 x, 分类器 $F(\cdot)$, 路标图片对应的真实标签 y_{true}

输出: 对抗样本 x_{adv}

1. 初始化伪二维码 M
2. **while** $F_t(x)=y_{true}$ **do**
3. // 判断输入的 x 经过 DNNs 分类后的输出标签是否与 y_{true} 一致, 若一致方可进行接下来的实验。
4. $G(i,j)\leftarrow F_t(x)$ //通过 DNNs 分类得到特征图 $G(i,j)$, $s.t.(i,j)\in\{0,1,2,\cdots,299\}$
5. $E(a,b)=\max(G(a_i,b_i))$ //挑选出特征图中对 DNNs 分类得到 y_{true} 影响最大的点 $E(a,b)$
6. 将伪二维码补丁的中心坐标点与 $E(a,b)$ 重合, 得到初始对抗样本 x'
7. **while** $F_t(x')=y_{true}$ **do**
8. // 初始对抗样本 x' 攻击失败, 进行优化
9. $x'_{N+1}=Clip_{x,\varepsilon}\left\{x'+\alpha\cdot\mathrm{sign}\left\{\nabla_t\left[l(F(x+t(M\cdot\varepsilon)),y^*)\right]\right\}\right\}$ //优化得到新的对抗样本 x'_{N+1}

10. $x' \leftarrow x'_{N+1}$ //将 x'_{N+1} 重新命名为 x'
11. **end while**
12. **if** $F_t(x') \neq y$ **then**
13. **return** $x_{adv} \leftarrow x'$ //攻击成功, 得到对抗样本 x_{adv}
14. **end while**
15. **return** x_{adv}

4 实验评估

4.1 实验设置

我们采用 VGG-16 模型[19]作为本次实验分类器, VGG-16 有 13 个卷积层(被 5 个 max-pooling 层分割)和 3 个全连接层。所有卷积核的大小都是 3×3, 步长为 1, 采用最大池化, 最后是一个输出层, 模型的输入大小为 224×224。实验数据由德国道路交通数据集[20](German traffic sign recognition benchmark, G-TSRB)和采集的中国道路交通标志部分图像组成了 42693 张的训练集 D_{train} 和 5200 张图片的测试集 D_{test}。其中德国交通数据集有 43 个类, 包含训练样本 39209 张, 测试样本 4300 张。我们对自行采集的道路交通标志图像进行了剪裁、筛选后, 得到 9 个类, 包含训练样本 3484 张, 测试样本 900 张。实验模型利用公开的 VGG-16 初始权重[21]进行训练, 最终得到训练集准确率为 99.10%, 测试集准确率为 93.52%。

图 4 实验中所使用的训练数据集

Figure 4 Training data set used in the experiment

4.2 评估指标

本文在数字空间和物理空间下分别进行了实验, 数字空间中的实验主要分为有目标攻击、无目标攻击两个部分。其中, 无目标攻击成功率是伪二维码补丁攻击成功的图片数与所有测试图片的总数的比值。有目标攻击成功率是伪二维码补丁攻击到指定类别的图片数与所有测试图片总数的比值。物理空间中的实验是以无目标攻击为例。文献[14]认为目前还没有评估物理对抗扰动的标准化方法, 所以我们参考[14]后在物理空间中不同的距离和角度下进行实验。

数字空间实验 无目标攻击成功率 acc_{UTA}:

$$acc_{UTA} = \frac{\sum_{x \in D_N} {}^{I}\{F(x') \neq y_{true} \wedge F(x) = y_{true}\}}{\sum_{x \in D_N} {}^{I}\{F(x) = y_{true}\}} \tag{10}$$

其中, D_N 是一组用于测试攻击成功率的路标数据集, ${}^{I}\{\cdot\}$ 是指示函数, x 是正常样本, x' 是对抗样本, y_{true} 是正常样本 x 对应的真实标签。

有目标攻击成功率 acc_{TA}:

$$acc_{TA} = \frac{\sum_{x \in D_N} {}^{I}\{F(x') = y^* \wedge F(x) = y_{true}\}}{\sum_{x \in D_N} {}^{I}\{F(x) = y_{true}\}} \tag{11}$$

其中, y^* 是对抗样本 x' 对应的指定类别标签。

物理空间实验 我们在不同的距离 $d \in D$ 和不同的角度 $g \in G$ 处获得图像 x, $c^{(d,g)}$ 表示从距离 d 和角度 g 拍摄的图像。x' 表示对抗样本。相机保持垂直拍摄且高度恒定。攻击成功率 acc_{LAB}:

$$acc_{LAB} = \frac{\sum_{x \in D_N} {}^{I}\{F(x') = y^* \wedge F(x^{(d,g)}) = y_{true}\}}{\sum_{x \in D_N} {}^{I}\{F(x^{(d,g)}) = y_{true}\}} \tag{12}$$

4.3 实验结果

4.3.1 数字空间下的攻击实验

我们在 D_{test} 中筛选出被 DNNs 正确分类的道路交通标志图像作为攻击测试样本, 如果图片在没有叠加扰动的情况下就被分类错误, 那么也就失去了制作对抗样本的意义。实验根据二维码的生成规则设计了三种规格的伪二维码 M: 大小分别为 21×21、25×25、29×29。为了更加贴近实际场景, 本节设计了两个实验条件进行对比: (1)三通道下的黑白伪二维码(RGB 三个通道的值只能同时为 0 或 255); (2)三通道下的彩色伪二维码(RGB 三通道的值在 0~255 像素之间); (3)限制三个定位符颜色为黑色, 目的是为了保证攻击补丁在攻击成功的基础上保持较高的伪装性。

实验从测试数据集中随机抽取 11 个类别进行测试。表 1 得到的数据是使用黑白的伪二维码补丁在 δ=20 时得到的攻击成功率。第一列表示不同的类别, 第二至四列是不同大小的伪二维码补丁攻击的成功

率。同样在彩色伪二维码补丁下重复了上述实验步骤, 结果如表 2 所示。

表 1　黑白的伪二维码补丁在 $\delta = 20$ 时的攻击成功率

Table 1　Attack success rate of black and white pseudo QR code patch at $\delta = 20$

类别	21×21/%	25×25/%	29×29/%
限速 30	49.09	63.64	76.36
限速 50	90.91	96.36	100.00
限速 70	89.09	96.36	98.18
限速 90	87.27	89.09	89.09
STOP	72.73	78.18	83.64
直行	78.18	83.64	85.45
靠左侧车道行驶	80.00	90.91	98.18
靠右侧车道行驶	69.09	72.73	94.55
向右急转	83.64	92.73	96.36
向左急转	94.55	96.36	96.36
注意行人	90.91	92.73	94.55

表 2　彩色的伪二维码补丁在 $\delta = 20$ 时的攻击成功率

Table 2　Attack success rate of color pseudo QR code patch at $\delta = 20$

类别	21×21/%	25×25/%	29×29/%
限速 30	54.55	83.64	98.18
限速 50	94.55	100.00	100.00
限速 70	94.55	100.00	100.00
限速 90	89.09	94.55	96.36
STOP	85.45	94.55	80.00
直行	83.64	92.73	92.73
靠左侧车道行驶	85.45	90.91	98.18
靠右侧车道行驶	74.55	83.64	96.36
向右急转	85.45	94.55	100.00
向左急转	96.36	96.36	98.18
注意行人	92.73	94.55	96.36

实验发现, 本文提出的二维码补丁攻击算法生成的对抗样本非常有效, 其攻击成功率可以达到 100%。但来自不同类别的道路交通标志在大小相同的伪二维码补丁扰动下, 攻击成功率有较大差异, 主要是因为某些类别的道路交通标志在 DNNs 分类中更容易受到攻击。

对于来自相同类别的道路交通标志, 面积更大的伪二维码补丁扰动会造成更大的影响。从表 1 中可以看出, 随着伪二维码补丁面积的增大, 攻击成功率也随随之增加。例如在限速 30 的类别中, 随着伪二维码补丁面积由 21×21 增大到 29×29 , 攻击成功率由 49.09%上升至 76.36%。但本文中的用于攻击的伪二维码攻击补丁的面积占整个路标的面积平均只有 0.95%, 大约仅为文献[14]中攻击补丁面积的 1/4。

从表 1, 表 2 中可以发现在相同的类别和扰动补丁面积下, 彩色伪二维码的攻击成功率会更高。比如在限速 30 的类别在 21×21 大小补丁的攻击下, 攻击成功率从黑白伪二维码补丁的 49.09%升高到彩色伪二维码补丁的 54.55%; 同样在 29×29 的伪二维码补丁中, 攻击成功率直接从 76.36%增加至 98.18%, 这是因为在相同大小的补丁中, 彩色补丁会携带更丰富的像素信息, 影响 DNNs 对于图片的特征提取, 从而有更大的概率将对抗样本误分类。

为了研究不同的迭代次数对伪二维码补丁攻击成功率的影响。实验以向左急转弯交通标志图像为例, 分析了在不同大小、颜色下的伪二维码补丁攻击的实验结果。如图 6 所示:

图 5　通过特征图添加伪二维码补丁得到初始化的对抗样本示例

Figure 5　Example of initial confrontation by adding pseudo QR code patches to the feature map

如图 6(a~f)可以看出, 在不同伪二维码面积大小、δ 和两种不同的伪二维码攻击中, 攻击成功率存在较大差异, 如(a)(c)(e)所示, 在 21×21 大小, 迭代次数均为 5 次的条件下, δ 越大, 攻击成功率就越高。在 δ 一定时, 伪二维码补丁面积越大攻击成功率也越高。例如(a)中, 当迭代次数为 5 次时, 21×21 大小的伪二维码补丁攻击成功率不到 40%, 但是 29×29 的攻击成功率达到了 65%。

观察图 6(e)(f)可以发现, 在 $\delta = 30$ 和攻击补丁大小均为 29×29 的情况下, 彩色伪二维码在第 6 次迭代时攻击成功率就已经达到了 100%, 而黑白伪二维码需要迭代 9 次, 这说明在相同情况下, 与黑白类伪二维码攻击相比, 彩色伪二维码攻击成功率要高出

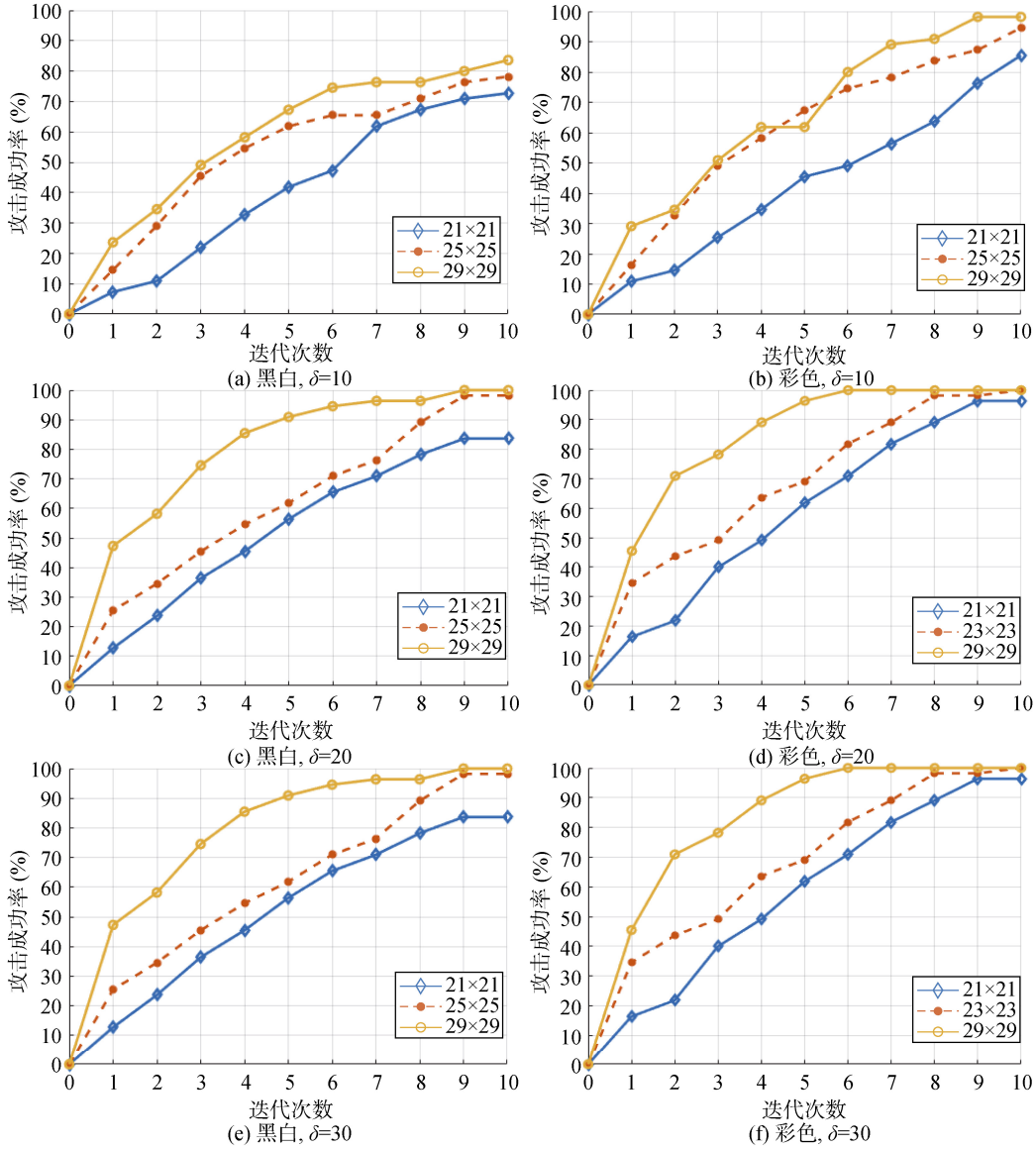

图 6　不同 δ 下, 黑白和彩色伪二维码在不同迭代次数下的攻击成功率

Figure 6　Attack success rate of black-and-white and color pseudo QR codes in different iterations under different δ

很多。因此, 在实际攻击中, 攻击者可优先使用彩色伪二维码进行攻击, 这不仅可以在短时间内生成攻击样本, 而且确保了较高的攻击成功率。

由此可见, 伪二维码面积、迭代次数、类型均会影响到攻击成功率。在本次实验中, 29×29 大小的伪二维码补丁面积占比依旧低于文献[14]中攻击补丁 4.8%占比, 实验证明我们的方法能够以较小代价生成隐蔽性强且攻击成功率高的对抗补丁。

本文方法同样适用于有目标攻击, 以向左急转交通标志图像为例设定实验方案: 输入正常的样本 x 得到该样本的显著性图, 本实验中选择对 DNNs 正确分类影响最大的前 20 个点作为扰动补丁添加的位置, 并在 $\delta=20$ 的情况下限定每个点的迭代上限为 10 次, 直至攻击成功得到对抗样本或者在前 20 个点中每个点达到迭代上限后依旧无法获得对抗样本而攻击失败。实验结果如表 3 所示。

通过表 3 可以观察到, 向左急转弯交通标志被误分类到其他类的成功率是不同的, 这说明有些类更容易被攻击。如表中彩色类别 29×29 的伪二维码补丁面积下, 分类到 STOP 类的成功率只有 45.45%, 而分类到向右急转类的成功率高达 81.82%。这说明在相同的情况下, DNNs 对 STOP 类的鲁棒性较高, 对向右急转类的鲁棒性较低。随着伪二维码补丁面积的不断增大, 有目标攻击的成功率总体上呈现上升趋势, 在同一类别中彩色伪二维码补丁有目标攻击成功率要高于黑白类伪二维码补丁。同时, 由实验可知, 在数字空间下, 有目标攻击比无目标攻击成功的难度更大。Papernot 等人[22]认为这可能是因为

DNNs 对某些类别的输入更难被错误分类, 从而导致有目标攻击时更加困难。

表 3 数字空间下 $\delta = 20$ 的有目标攻击成功率

Table 3 Success rate of targeted attack with $\delta = 20$ in digital space

	大小	向右急转 (%)	注意行人 (%)	STOP (%)	T 型左前交叉 (%)	T 型左右交叉 (%)
黑白	21×21	18.18	9.09	9.09	18.18	27.27
	25×25	36.36	18.18	27.27	36.36	27.27
	29×29	54.55	36.36	36.36	45.45	36.36
彩色	21×21	36.36	9.09	18.18	27.27	36.36
	25×25	54.55	27.27	27.27	45.45	54.55
	29×29	81.82	54.55	45.45	63.64	72.73

为了测试在 VGG-16 分类器中产生的对抗样本是否具有可转移性, 我们重新训练了一个新的分类器 VGG-19[19], 并将在 VGG-16 中攻击成功的对抗样本在 $\delta = 20$ 的情况下对 VGG-19 进行测试, 实验结果如表 4 所示。

通过表 4 可以看出, 针对 VGG-16 的对抗样本在 VGG-19 的模型分类中仍然可以攻击成功。虽然攻击成功率在分类器进行替换之后有一定的降低, 但是总体上来说, 随着伪二维码面积增加, 攻击的成功率也在逐渐提高。因此实验中利用式(14)产生的对抗样本具有一定的攻击转移性。

表 4 数字空间下 $\delta = 20$ 时对 VGG-19 分类器的攻击成功率

Table 4 Attack success rate of VGG-19 classifier with $\delta = 20$ in digital space

	大小	向右急转 (%)	注意行人 (%)	STOP (%)	T 型左前交叉 (%)	T 型左右交叉 (%)
黑白	21×21	9.09	3.64	1.82	7.27	21.82
	25×25	32.73	14.55	27.27	21.82	16.36
	29×29	50.91	23.64	36.36	23.64	18.18
彩色	21×21	25.45	10.91	18.18	16.36	21.82
	25×25	41.82	12.73	27.27	23.64	30.91
	29×29	51.18	34.55	45.45	47.27	41.82

4.3.2 物理空间下的攻击实验

4.3.1 中的实验均在数字空间下进行, 我们以向左急转弯道路交通标志为例设计了物理空间下的实验, 本节将在数字空间下攻击成功的伪二维码补丁直接进行打印后在物理空间中进行测试。实验中, 拍摄角度均为 90°正面拍摄, 并在不同的距离下每组拍摄得到的 55 张照片来测试, 本节以式(16)来衡量攻击成功率, 实验发现黑白的贴纸攻击成功率最高为 83.64%。而彩色贴纸攻击成功率最高达到了 90.91%。我们将攻击成功之后的置信度绘制成图 7 和图 8。

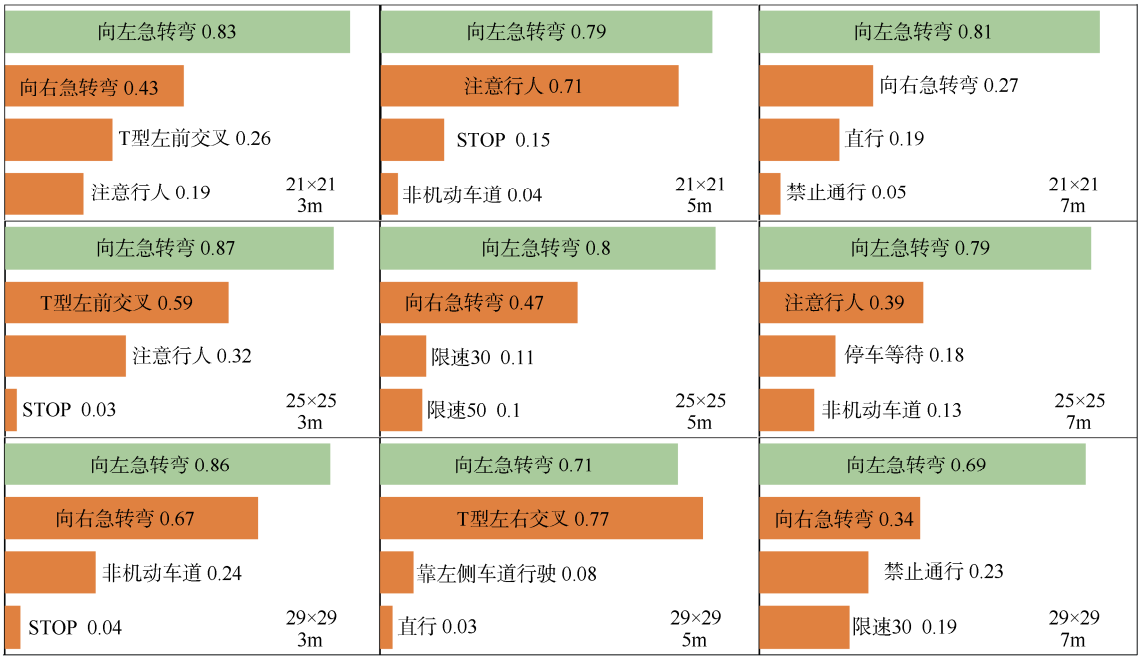

图 7 黑白的伪二维码补丁在不同距离下攻击成功之后前三类的置信度

Figure 7 The confidence of the first three types of black and white pseudo QR code patches after successful attacks at different distances

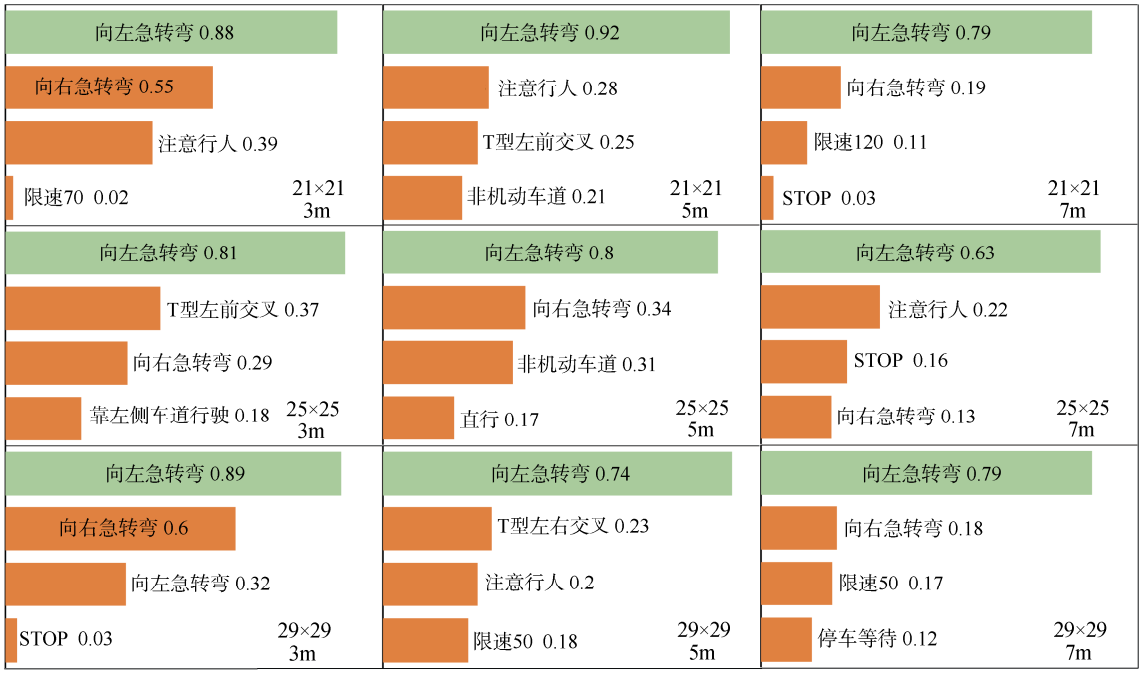

图 8　彩色的伪二维码补丁在不同距离下攻击成功之后前三类的置信度

Figure 8　The confidence of the first three types of color pseudo QR code patches after successful attack at different distances

图 7 和图 8 是向左急转弯道路交通标志在不同伪二维码补丁大小和距离下攻击成功后得到新分类的平均置信度, 其中限定相机在恒定高度下保持正面角度拍摄图像。图中的浅绿色柱体代表正常样本在物理空间下分类正确之后的平均置信度, 橙色柱体代表添加二维码补丁后攻击成功的平均置信度最大的前三类。

由 4.3.1 以及 4.3.2 中的实验对比可知, 物理空间下向左急转弯有很大的概率被判断为向右急转, 主要是因为两种类别的图像存在较强的关联性, DNNs 从中抽取了的特征信息较为相似。

同样地, 我们进行了户外现场测试实验, 如图 9 所示, 为了更好的模拟无人驾驶的行驶行为, 测试中分别设计了不同距离、不同角度下的攻击实验。结果如表 5 所示:

表 5　真实场景下攻击成功后的置信度

Table 5　Confidence after successful attack in real scene

		原图	中号伪二维码	大号伪二维码
黑白	0m 0°	向左急转弯 0.81	向右急转弯 0.52	注意行人 0.66
	10m 15°	向左急转弯 0.78	STOP 0.67	停车等待 0.53
	10m 30°	向左急转弯 0.72	T 型左右交叉 0.42	向右急转弯 0.41
	15m 0°	向左急转弯 0.79	限速 120 0.39	靠左侧车道行驶 0.39
	15m 15°	向左急转弯 0.77	向右急转弯 0.42	禁止通行 0.45
	15m 30°	向左急转弯 0.68	注意行人 0.34	向右急转弯 0.52
彩色	10m 0°	向左急转弯 0.87	注意行人 0.46	T 型左右交叉 0.38
	10m 15°	向左急转弯 0.80	禁止通行 0.19	向右急转弯 0.44
	10m 30°	向左急转弯 0.77	靠左侧车道行驶 0.23	向右急转弯 0.56
	15m 0°	向左急转弯 0.84	限速 30 0.36	STOP 0.39
	15m 15°	向左急转弯 0.82	向右急转弯 0.32	T 型左右交叉 0.67
	15m 30°	向左急转弯 0.76	停车等待 0.45	向右急转弯 0.52

表 5 中第二列表示在不同距离和角度下进行的实验设置, 第三列原图为分类正确后的向左急转弯道路交通标志的平均置信度。第一行中的中号和大号分别代表两种大小的伪二维码, 其尺寸是按照现

场真实道路交通标志面积的大小设计的, 面积占比依旧符合实验的设计原则。第四列和第五列的数据是攻击成功后的类别的平均置信度。在现场测试中, 黑白和彩色伪二维码攻击成功率分别为 78.92%和88.37%。而且在较大角度和距离变化情况下, 通过式(14)生成的对抗样本依旧具有鲁棒性。

由 4.3.1 以及 4.3.2 中的实验对比可知, 物理空间下的攻击成功率相比数字空间下较弱, 主要是因为数字空间的对抗样本在转移到物理空间中时, 可能由于相机本身传感器缺陷、印刷不完美等原因导致的。但排除这些原因, 我们依旧得到了较好的攻击结果。

图 9 户外现场测试
Figure 9 Outdoor field test

5 结论

近年来, 随着硬件计算能力的提升, 训练算法的改进和数据的不断丰富, 深度神经网络得到了迅速发展, 并广泛应用于各种领域。图像识别领域, 各种深度神经网络模型不断提出, 从基础的 Lenet-5[23]到越来越深层的 VGG、GoogLeNet[24]等, 深度神经网络的能力不断得到提高。然而, 深度神经网络快速发展的背后也存在着诸多问题, 其中对抗样本的存在让人们开始关注深度神经的安全性问题。

在以往的工作中, 研究者更关注在数字空间中存在的对抗样本, 通过对图像样本进行优化来寻求在细微的扰动下的对抗样本, 但是随着无人驾驶技术的普及, 物理对抗样本开始引起广泛关注, 在本文中, 我们以伪二维码作为攻击补丁, 成功的在数字空间和物理空间中攻击了分类器, 并且取得了不错的攻击效果, 本文认为, 深度神经网络在快速发展的同时, 其安全问题应该得到进一步的重视。

参考文献

[1] Yin X, Liu X M. Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition[EB/OL]. 2017: arXiv:1702.04710 [cs.CV]. https://arxiv.org/abs/1702.04710.

[2] Li H, Wang P, You M Y, et al. Reading Car License Plates Using Deep Neural Networks[J]. *Image and Vision Computing*, 2018, 72: 14-23.

[3] Kumar A, Irsoy O, Ondruska P, et al. Ask me Anything: Dynamic Memory Networks for Natural Language Processing[EB/OL]. 2015: arXiv:1506.07285[cs.CL]. https://arxiv.org/abs/1506.07285.

[4] Szegedy C, Zaremba W. Intriguing properties of neural networks[C]. *Computer Vision and Pattern Recognition*, 2014: 57-92.

[5] https://keenlab.tencent.com/zh/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/.

[6] Papernot N, McDaniel P, Goodfellow I. Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples[EB/OL]. 2016: arXiv:1605.07277[cs.CR]. https://arxiv.org/abs/1605.07277.

[7] Goodfellow J, Shlens J, Szegedy C, Explaining and harnessing adversarial examples. 2014: arXiv preprint arXiv:1412.6572.

[8] Xie C, Zhang Z, Zhou Y. Improving transferability of adversarial examples with input diversity[C]. *Computer Vision and Pattern Recognition*, 2019: 2730-2739.

[9] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[EB/OL]. 2015: arXiv:1511.07528 [cs.CR]. https://arxiv.org/abs/1511.07528.

[10] Carlini N, Wagner D. Towards evaluating the ro bustness of neural networks[C]. *IEEE Symposium on Security and Privacy (SP)*, 2017: 39-57.

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

[12] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[EB/OL]. 2013: arXiv:1312.6034[cs.CV]. https://arxiv.org/abs/1312.6034.

[13] Carlini N, Wagner D. Adversarial Examples are not Easily Detected: Bypassing Ten Detection Methods[C]. *the 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 3-14.

[14] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Models[EB/OL]. 2017: arXiv:1707.08945[cs.CR]. https://arxiv.org/abs/1707.08945.

[15] Sitawarin C, Bhagoji A N, Mosenia A, et al. Rogue Signs: Deceiving Traffic Sign Recognition with Malicious Ads and Logos[EB/OL]. 2018: arXiv:1801.02780[cs.CR]. https://arxiv.org/abs/1801.02780.

[16] Kingma D P, Ba J. Adam: A Method for Stochastic

Optimization[EB/OL]. 2014: arXiv:1412.6980[cs.LG]. https://arxiv.org/abs/1412.6980.

[17] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[EB/OL]. 2016: arXiv:1608.04644[cs.CR]. https://arxiv.org/abs/1608.04644.

[18] Liu Y, Yang J, Liu M. Recognition of QR Code with mobile phones[C]. *Chinese Control And Decision Conference*, 2008: 203-206.

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

[20] Stallkamp J, Schlipsing M, Salmen J, et al. Man Vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition[J]. *Neural Networks*, 2012, 32: 323-332.

[21] VGG-16 initial weight, https://pan.baidu.com/s/1UfHxzHwHC-hXNr0WXzSlQNg.

[22] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016: 372-387.

[23] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[24] Szegedy C, Liu W, Jia Y. Going deeper with convolutions[C]. *Computer Vision and Pattern Recognition*, 2015: 1-9.

钱亚冠(1976-) 男, 浙江嵊州人, 于 2014 年在浙江大学获得博士, 浙江科技学院副教授, 硕士生导师, 主要研究方向: 机器学习与大数据处理、对抗性机器学习、目标识别与跟踪。Email: qianyaguan@zust.edu.cn

顾钊铨(1989-) 男, 分别于2011, 2015年获得清华大学计算机科学学士、博士学位。现为广州大学网络空间高级技术学院(CIAT)的教授。研究方向：无线网络、分布式计算、大数据分析和人工智能安全。Email: zqgu@gzhu.edu.cn

刘新伟(1996-) 男, 河南信阳人, 硕士研究生, 就读于浙江科技学院, 研究方向: 对抗机器学习。Email: crow_821@163.com

王滨(1978-) 男, 博士, 教授级高级工程师, 杭州海康威视数字技术有限公司副总裁。研究方向: 网络与信息安全、人工智能安全性、物联网安全性等。 Email: wbin2006@gmail.com

潘俊(1978-) 男, 于 2011 年在浙江大学获得计算机专业博士学位, 主要研究方向为自然语言处理、数据挖掘。Email: panjun@zust.edu.cn

张锡敏(1996-) 于 2018 年获得徐州大学学士学位。现为浙江科技学院大数据学院研究生。研究方向主要为模式识别和机器学习安全性。Email: 1289486793@qq.com