软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(9):2785−2801 [doi: 10.13328/j.cnki.jos.005945]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于 PSO 的路牌识别模型黑盒对抗攻击方法*

陈晋音, 陈治清, 郑海斌, 沈诗婧, 苏蒙蒙

(浙江工业大学 信息工程学院,浙江 杭州 310023)

通讯作者: 陈晋音, E-mail: chenjinyin@zjut.edu.cn

摘 要: 随着深度学习在计算机视觉领域的广泛应用,人脸认证、车牌识别、路牌识别等也随之呈现商业化应用趋势.因此,针对深度学习模型的安全性研究至关重要.已有的研究发现:深度学习模型易受精心制作的包含微小扰动的对抗样本攻击,输出完全错误的识别结果.针对深度模型的对抗攻击是致命的,但同时也能帮助研究人员发现模型漏洞,并采取进一步改进措施.基于该思想,针对自动驾驶场景中的基于深度学习的路牌识别模型,提出一种基于粒子群优化的黑盒物理攻击方法(black-box physical attack via PSO,简称 BPA-PSO).BPA-PSO 在未知模型结构的前提下,不仅可以实现对深度模型的黑盒攻击,还能使得实际物理场景中的路牌识别模型失效.通过在电子空间的数字图像场景、物理空间的实验室及户外路况等场景下的大量实验,验证了所提出的 BPA-PSO 算法的攻击有效性,可发现模型漏洞,进一步提高深度学习的应用安全性.最后,对 BPA-PSO 算法存在的问题进行分析,对未来的研究可能面临的挑战进行了展望.

关键词: 自动驾驶;对抗性攻击;路牌识别;黑盒物理攻击;粒子群优化

中图法分类号: TP18

中文引用格式: 陈晋音,陈治清,郑海斌,沈诗婧,苏蒙蒙.基于 PSO 的路牌识别模型黑盒对抗攻击方法.软件学报, 2020,31(9):2785−2801. http://www.jos.org.cn/1000-9825/5945.htm

英文引用格式: Chen JY, Chen ZQ, Zheng HB, Shen SJ, Su MM. Black-box physical attack against road sign recognition model via PSO. Ruan Jian Xue Bao/Journal of Software, 2020,31(9):2785−2801 (in Chinese). http://www.jos.org.cn/1000-9825/5945.htm

Black-box Adversarial Attack Against Road Sign Recognition Model via PSO

CHEN Jin-Yin, CHEN Zhi-Qing, ZHENG Hai-Bin, SHEN Shi-Jing, SU Meng-Meng

(School of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: With the wider application of deep learning in the field of computer vision, face authentication, license plate recognition, and road sign recognition have also presented commercial application trends. Therefore, research on the security of deep learning models is of great importance. Previous studies have found that deep learning models are vulnerable to carefully crafted adversarial examples that contains small perturbations, leading completely incorrect recognition results. Adversarial attacks against deep learning models are fatal, but they can also help researchers find vulnerabilities of models and make further improvements. Motivated by that, this study proposes a black box physical attack method based on particle swarm optimization (BPA-PSO) for deep learning road sign recognition model in scenario of autonomous vehicles. Under the premise of unknown model structure, BPA-PSO can not only realize the black box attack on

* 基金项目: 浙江省自然科学基金(LY19F020025); 国家重点研发计划(2018AAA0100800); 宁波市“科技创新 2025”重大专项(2018B10063); 浙江省认知医疗工程技术研究中心(2018KFJJ07)

Foundation item: Zhejiang Provincial Natural Science Foundation of China (LY19F020025); National Key Research and Development Program of China (2018AAA0100800); Major Special Funding for “Science and Technology Innovation 2025” in Ningbo (2018B10063); Engineering Research Center of Cognitive Healthcare of Zhejiang Province (2018KFJJ07)

本文由“智能嵌入式系统”专题特约编辑王泉教授、吴中海教授、陈仪香教授、苗启广教授推荐.

收稿时间: 2019-07-03; 修改时间: 2019-08-18; 采用时间: 2019-11-02; jos 在线出版时间: 2020-01-13

CNKI 网络优先出版: 2020-01-14 11:27:06, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.1126.025.html

deep learning models, but also invalidate the road sign recognition models in the physical scenario. The attack effectiveness of BPA-PSO algorithm is verified through a large number of experiments in the digital images of electronic space, laboratory environment, and outdoor road conditions. Besides, the abilities of discovering models' vulnerabilities and further improving the application security of deep learning are also demonstrated. Finally, the problems existing in the BPA-PSO algorithm are analyzed and possible challenges of future research are proposed.

Key words: autopilot; adversarial attack; road recognition; black-box physical attack; particle swarm optimization

深度学习凭借其强大的特征提取与拟合能力而被广泛应用于各个领域,如自动驾驶[1]、人脸识别[2]、语音识别[3]、恶意软件检测[4]、推荐系统[5]、生物信息[6]、城市管理[7]、目标检测与识别[8−14]等.其中,自动驾驶技术的日趋成熟引起了研究人员的广泛关注,其涵盖了图像处理、语音识别、激光雷达、GPS 定位、自动路径规划等大量前沿技术.而无人汽车的高级控制系统依赖基于深度学习的图像识别与目标检测等实现环境的感知,其中,基于深度学习(例如 GTSRB-CNN[15])的路牌识别是主要技术之一.

然而,最新的研究发现,深度学习容易受到对抗样本攻击[16],即:通过在正常良性样本中添加精心设计的微小扰动得到的对抗样本,可使原本分类准确率接近 99%的深度学习模型完全失效,且添加的扰动肉眼不可见.因此,对抗样本具有较强的迷惑性和危害性.Szegedy 等人[17]首次证明了:通过在输入数据中添加小规模精心制作的扰动,能够使卷积神经网络做出错误决策.此后出现了更多的针对深度模型的对抗攻击方法.根据深度模型的透明程度,可以分为白盒攻击和黑盒攻击:白盒攻击如 FGSM[18]、C&W[19]、DeepFool[20]、通用对抗扰动攻击[21]、单像素攻击[22]等;黑盒攻击如 Boundary[23],ZOO[24],POBA-GA[25].这些攻击方法计算得到的对抗扰动虽然是不明显的,甚至是肉眼不可见的,但能够导致深度学习模型失效.对抗攻击不仅发生在数字虚拟空间,也出现在现实物理空间中.在物理世界,Kurakin 等人[26]通过手机摄像头识别打印的对抗样本时出现错误分类;Sharif 等人[27]研制出了一副带有对抗扰动的"眼镜",可以让佩戴者躲过人脸识别系统或者被误识为另一个人;马玉琨等人[28]提出了一种面向人脸活体检测的对抗样本生成算法.

在物理场景的攻击中,通常需要面对这些挑战:(1) 物理场景的背景环境是多变的,无法通过控制背景进行攻击;(2) 光线、距离和角度的不同容易引起对抗扰动的攻击失效;(3) 扰动过小可能使图像传感器无法有效捕捉,太大则容易引起人眼的警觉.

自动驾驶车辆的路牌识别系统,其安全性和可靠性对汽车行驶过程中做出正确的决策具有重要影响.当攻击者在真实的路牌上添加对抗扰动,并成功攻击路牌识别系统时,可能会造成难以想象的灾难.本文针对物理世界中常用的基于深度学习的路牌识别系统展开攻击,设计基于粒子群优化的黑盒物理攻击方法(black-box physical attack via PSO,简称 BPA-PSO),攻击者可以操控被攻击对象的物理外表,如在路牌上添加一些不易引起人类警觉的海报或贴纸来欺骗自动驾驶车辆的路牌识别系统.BPA-PSO 算法在不知道目标模型结构和参数等细节的前提下,通过迭代优化得到在物理世界实现有效攻击的对抗样本,该方法能够克服真实路牌识别场景中的光线、角度和距离等因素的影响,发现基于深度学习的路牌识别系统中存在的安全漏洞.

根据攻击者希望实现的攻击目标的不同,可以分为有目标攻击和无目标攻击.在无目标攻击中,攻击者的目的是使某一路牌不能被正确识别或者被识别成其他任意一种路牌.例如:攻击"紧急转弯"标识后,被路牌识别系统错误识别为任意其他标志,使得车辆遇到急转弯时发生意外.在有目标攻击中,攻击者试图在某一路牌上添加海报或贴纸使该路牌标志被识别为指定的另一种路牌标志,这类攻击往往带有更大的危害性.例如:当一个区域内被攻击的路牌标志数量较多,甚至可以构成一个系统时,该区域的交通很容易发生瘫痪,造成较大的损失.BPA-PSO 算法通过调整适应度函数的优化目标,能够同时实现有目标攻击和无目标攻击.

本文专注于研究基于进化计算的路牌识别攻击有两个重要的原因:首先,这种攻击属于黑盒攻击,符合物理世界中无法获得模型细节的真实情况,并且能够产生对环境具有较强鲁棒性的对抗扰动;其次,对于自动驾驶车辆来说,当它的路牌识别系统受到攻击时,所引起的后果往往是灾难性的,因此,研究这种攻击有助于学习如何进行防御.图 1 展示了 BPA-PSO 算法得到的能够成功攻击物理世界中路牌识别系统的对抗样本,从左到右分别是禁止鸣笛的正常路牌图像、添加在正常路牌上的对抗扰动、在电子空间中添加扰动后的对抗路牌样本、在

物理世界中添加扰动后打印并拍摄的对抗路牌样本.其中:正常路牌图像在物理世界中打印后能够被正确识别,对抗路牌图像在物理世界中打印后被路牌识别系统错误识别为“限速 40km/h”.本文的工作将有助于理解物理世界的自动驾驶车辆中基于深度学习的图像识别模型,检测已有识别模型的安全漏洞,为进一步提高深度学习模型鲁棒性的研究工作提供帮助.

Fig.1 Example of an adversarial image that successfully attacks against road sign recognition system

图 1 成功攻击路牌识别系统的对抗样本图举例

本文的主要贡献如下.

(1) 设计了基于粒子群优化的黑盒攻击方法,不需要了解模型结构和参数等细节,能够仅通过模型输出类标和最高置信度的信息实现有效攻击,符合实际应用场景,同时具有较好的迁移性;

(2) 生成的对抗扰动在物理场景下攻击有效且不易引起人类警觉,通过在电子空间中模拟物理场景,将对抗样本旋转、缩放和光影变化后的攻击效果作为优化目标的一部分,提高了物理攻击的稳定性和可靠性;

(3) 本文建立了一个新的中国路牌数据集和对抗路牌数据集,通过大量实验,验证了 BPA-PSO 算法在电子空间和物理空间中针对基于深度学习的路牌识别系统的攻击有效性.

本文第 1 节对目前针对深度学习模型的对抗攻防研究进行总结,包括主流的白盒攻击方法、黑盒攻击方法以及目前主流的优化算法介绍.第 2 节对本文提出的 BPA-PSO 算法进行介绍,详细说明如何在路牌标志的局部区域上添加海报或贴纸来欺骗自动驾驶车辆的路牌识别系统,以及如何提高对抗攻击的物理稳定性和可靠性.第 3 节对实验设计和结果分析进行阐述,说明了 BPA-PSO 算法在多种物理场景中的攻击有效性和可靠性.第 4 节和第 5 节分别对本文进行总结,对未来的工作和挑战进行展望.

1 相关工作

1.1 白盒攻击方法介绍

攻击者能够获得机器学习所使用的模型结构以及模型的参数,并利用它们产生对抗样本数据的攻击称为白盒攻击,在这种攻击过程中,更多地需要与模型的内部参数信息之间进行交互.本小节介绍了几种主流的白盒攻击方法.

(1) 快速梯度符号法(FGSM)

FGSM[18]是 Goodfellow 等人提出的生成对抗样本的一种简单算法,其主要思想是:计算深度神经网络模型梯度变化最大的方向,并在该方向上添加对抗性扰动,通过增加模型损失的方式,使得模型进行错误的分类.扰动计算公式如下:

$$\rho=\varepsilon\cdot sign(\nabla_x J(\theta,x,y)) \tag{1}$$

其中,$\nabla_x J(\cdot)$是在当前参数θ下,损失函数在原始图像附近计算得到的梯度;$sign(\cdot)$表示符号函数;y 是结果标签.

(2) C&W 攻击

C&W 攻击是由 Carlini 和 Wagner[19]提出的基于优化的攻击,通过限制对抗扰动的 0-范数、2-范数或者无穷范数,使扰动变得几乎不可察觉.因此,这类攻击的成功需要满足对抗样本与原图的差距越小越好.对抗样本使得模型分类错误,且错的那一类的置信度越高越好.实验证明,针对目标网络的防御几乎无法抵御这类白盒攻

击生成的对抗样本.

(3) JSMA 攻击

在主流的对抗攻击方法中,常用的是限制扰动的 2-范数或无穷范数来限制扰动使得人眼无法察觉.然而,Papernot 等人[29]提出了限制扰动 0-范数的方法也能够实现对抗攻击,并且这种方法只需修改图像中的几个像素点的值.该算法的主要思想是:一次只修改一个原始图像的像素,并通过网络层的输出梯度计算显著图来监视修改后对分类结果的影响.在显著图中,数值越大表示欺骗网络的可能性越高.该算法根据计算出的显著图像和当前图像,选择其中最有效的像素点进行修改从而欺骗网络.

(4) Houdini 攻击

Houdini 是由 Cisse 等人[30]提出的一种通过产生可以适应任务损失的对抗性样本来欺骗基于梯度的机器学习的算法.一般产生对抗样本的典型算法是采用网络损失函数的梯度来计算扰动.然而,有些任务损失函数往往不适合这种方法.例如:在语音识别中是根据字错误率来产生对抗性样本,而不是损失函数的梯度.Houdini 则是专门为这类任务提供产生对抗样本的方法.

(5) MI-FGSM

Dong 等人[31]提出了一种基于动量的迭代攻击算法来提升对抗性攻击能力,即 MI-FGSM.它将动量项添加到攻击的迭代过程中,这有利于加快收敛速度、使更新方向更加平稳,并在迭代期间能够从较差的局部最大值中逃脱,从而达到更好的攻击效果.

(6) 单像素攻击

Su 等人[22]在每幅图像中只改变一个像素点的情况下,使得 70.97%的图像在测试中成功地欺骗了 3 种不同的网络模型.而且网络错误分类时的平均置信度高达 97.47%.Su J 等人使用差分进化的概念来计算对于样本,通过对每个像素点进行修改生成子图,并与母图进行对比,根据选择标准保留攻击效果最好的子图像,从而实现对抗攻击.

1.2 黑盒攻击方法介绍

与白盒攻击相反,黑盒攻击是指在攻击者不知道目标模型信息的情况下生成对抗样本.在一些情况下可以假定攻击者对模型有一定的认识,但是绝对不知道目标模型的内部参数.因此,这种攻击往往更加符合实际.在本小节中,介绍了几种主流的黑盒攻击方法.

(1) UPSET 和 ANGRI 攻击

UPSET 和 ANGRI 是 Sarkar 等人[32]提出的两种黑盒攻击算法,其中:UPSET 可以作为特定目标类的目标攻击,在图像不可知时产生的对抗性扰动添加到任何图像上都可以使图像分类器将其识别成目标类别;ANGRI 则是作为特定图像的目标国际,其生成的是特定图像的扰动.在 MNIST 和 CIFAR10 数据集的实验中,这两种攻击方法都获得了高欺骗率.

(2) 零阶优化攻击(ZOO)

基于零阶优化的攻击是 Chen 等人[24]提出的一种有效的黑盒攻击,它是只访问模型的输入图像和输出的置信度分数,基于零阶优化,通过直接估计目标模型的梯度来生成对抗样本.这种攻击不需要训练替代模型,并避免了攻击可转移性的损失,是目前黑盒攻击中最有效的攻击方法之一.

(3) 边界攻击(boundary attack)

边界攻击是由 Brendel 等人[23]提出的一种基于决策的对抗攻击算法,它的主要思想是:从生成大的对抗性扰动开始,然后在保持对抗性扰动的同时,力求减少扰动.这种攻击几乎不需要超参数的调整,也不依赖于替代模型,只依赖模型的最终决策,并且这种攻击使得机器学习与真实世界的关联性更大,因为现实中我们很容易得到模型的决策结果而不是置信度分数或 log*it* 值.

1.3 群体智能优化算法介绍

本节主要介绍了几种常见的群体智能优化算法.

(1) 蚁群算法

蚁群算法是 Dorigo 等人[33]受到蚂蚁觅食现象的启发而提出的一种群体智能优化算法,属于随机搜索算法,其主要思想是人工模拟蚂蚁搜索食物的过程.蚁群算法特点是可以进行分布式计算、具有较强的鲁棒性以及容易同其他方法相结合.但与其他方法相比,该算法的复杂度较大、搜索时间较长,并且容易出现停滞现象.

(2) 粒子群优化算法(particle swarm optimization,简称 PSO)

粒子群优化算法是 Kennedy 等人[34]源于对鸟群觅食现象而提出的一种进化算法.在粒子群算法中,每个粒子能够记录下自己飞过的历史最优位置,粒子之间可以通过记忆信息共享实现群体的优化.粒子群算法是一种不需要梯度信息的全局优化算法.粒子群算法的参数较少、易于设置和调整、具有较快的收敛速度.但粒子群算法也存在易陷入局部最优的缺点,并且粒子群的初始解分布对全局最优解具有较大的影响.由于该算法出色的优化性能,本文采用粒子群算法来优化来攻击路牌识别模型.同时,针对该算法易陷入局部最优的缺点,本文通过改变初始解的生成方式来搜索全局近似最优解.

(3) 人工鱼群算法

人工鱼群算法是李晓磊等人[35]提出的一种群体智能优化算法.人工鱼群算法通过构造人工鱼来模仿鱼群的觅食行为、聚群行为、追尾行为和随机行为来实现寻优.该算法具有较快的收敛速度、对初值和参数选择不敏感、易于实现等优点.但对于较大规模的问题时求解困难,收敛较慢.

其他的群体智能优化算法还包括菌群算法[36]、蛙跳算法[37]、人工蜂群算法[38]等,它们都具有良好的性能和各自的特点.

2 基于 PSO 的路牌识别模型的攻击方法介绍

2.1 算法框架介绍

本文设计的基于 PSO 的路牌识别模型的攻击方法主要是通过 PSO 算法,在对路牌识别模型内部参数未知的情况下,通过迭代寻优来生成路牌识别模型的对抗样本.同时,利用图像处理技术在电子空间中模拟物理世界的噪声干扰来优化对抗扰动,提高对抗样本在物理世界的攻击鲁棒性.图 2 展示了生成物理世界攻击有效的对抗样本的过程:首先输入一张良性路牌图像,添加随机扰动得到多张对抗路牌图像作为初始解;然后使用 PSO 进行寻优.根据对抗路牌图像的攻击效果更新搜索方向,使得优化后对抗路牌图像的攻击成功率高、扰动的稳定性强以及扰动的不可见效果好.测试优化后的对抗路牌图像在物理世界中的攻击效果.

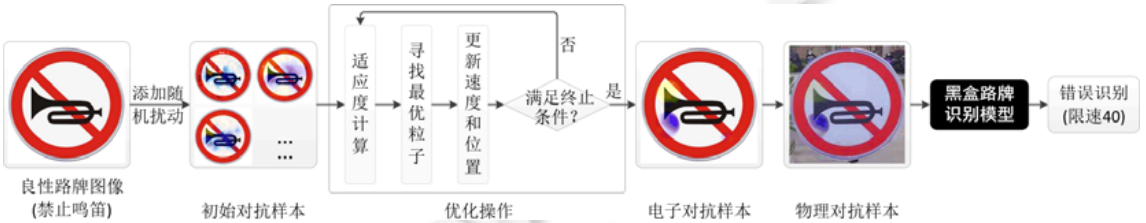


Fig.2 Framework of black-box physical attack against road sign recognition model via PSO

图 2 基于 BPA-PSO 的路牌识别模型的攻击方法框图

2.2 BPA-PSO算法具体步骤介绍

粒子的初始解对 PSO 算法的寻优结果具有重要影响.BPA-PSO 通过添加随机扰动获得对抗样本作为初始解,虽然大部分初始对抗路牌对于最终要攻击的目标路牌识别模型没有攻击效果,但是我们相信,其中部分特征与最终要攻击的目标路牌识别模型的对抗样本特征分布存在一致性.因此,BPA-PSO 算法通过寻优操作充分利用随机扰动的对抗性特征,将有用的特征保留,最终实现对目标路牌识别模型的攻击.

BPA-PSO 算法的具体描述如下.

(1) 粒子群算法初始化,将添加了随机扰动得到的每个对抗样本作为一个粒子,将每个对抗样本所有像素

点的 RGB 值作为粒子的位置矩阵 x_i,RGB 值的变化速度作为粒子的速度矩阵 v_i,以及当前迭代数 g,最大迭代数 G_k,当前惯性权重因子 $\omega^{(g)}$,第 i 个粒子的历史最优位置 p^{best_i},粒子种群发现的全局最优位置 g^{best_i};

(2) 计算粒子群的适应度值,对每个粒子进行随机图像变换,并计算每个粒子变换后的适应度值;

(3) 根据得到的粒子群的适应度值,比较历史最佳适应度,并更新每个粒子的历史最优位置 p^{best_i}、粒子群的全局最优位置 g^{best_i};

(4) 更新粒子群的速度 v_i 和位置 x_i.我们在粒子速度和位置的更新过程中采用了惯性因子,其值较大时全局搜索能力强,其值较小时局部搜索能力强,计算公式如下:

$$\omega^{(g)}=(\omega_{ini}-\omega_{end})(G_k-g)/G_k+\omega_{end} \tag{2}$$

$$\upsilon_i=\omega^{(g)}\times\upsilon_i+c_1\times rand(\cdot)\times(p^{best_i}-x_i)+c_2\times rand(\cdot)\times(g^{best_i}-x_i) \tag{3}$$

$$x_i=x_i+\upsilon_i \tag{4}$$

其中,ω_{ini} 为初始权重因子值,ω_{end} 为最终权重因子值,c_1 和 c_2 为初始化学习因子,$rand(\cdot)$为系统产生的介于(0,1)之间的随机数.

(5) 判断是否达到最大迭代数或全局最优解满足条件:若满足,则结束迭代,将搜索到的近似最优解作为最终的对抗样本;否则,返回步骤(2)继续迭代.

2.3 基于随机扰动的对抗路牌初始化

首先准备一组 50 张待攻击的良性路牌图像集合,集合中的图像要求在距离为 3m~6m、倾斜角为 0°时拍摄的清晰的路牌图像,并通过添加随机扰动获得初始路牌粒子.为了促进物理攻击的有效性,需要对扰动优化目标进行修改,在对扰动范数限制的基础上叠加了扰动平滑性制约(perturbation smoothness restriction,简称 PSR),计算公式如下:

$$f_{PSR}(\rho)=\frac{1}{n}\sum_{k,j}(\sum(\rho_{k,j}-\rho_{near(k,j)})^2)^{\frac{1}{2}} \tag{5}$$

其中,原始的扰动优化目标是最小化$\|\rho\|_2$,修改后的扰动优化目标是最小化$\|\rho\|_2+f_{PSR}(\rho)$,$\|\rho\|_2$ 表示对扰动的 2-范数限制;扰动$\rho=x^*-x$,x^*表示对抗路牌图像,x 表示良性路牌图像;$\rho_{i,j}$ 是扰动中坐标位置为(k,j)的扰动像素点的 RGB 三通道像素 $R_{k,j}$,$G_{k,j}$ 以及 $B_{k,j}$ 的平均值;$\rho_{k,j}=\frac{1}{3}(R_{k,j}+G_{k,j}+B_{k,j})$ 是扰动中与坐标位置为(k,j)相邻的所有扰动像素点的 RGB 三通道像素的平均值.

2.4 适应度函数设计

PSO 算法是以适应度函数为依据,通过比较种群每个个体的适应度值来进行搜索近似最优解.同样,适应度函数的设计将直接影响 BPA-PSO 算法搜索到的对抗扰动的性能.适应度函数包括 3 部分,分别是对抗性指标 $f_{adv}(x^*)$、扰动平滑度 $f_{PSR}(\rho)$以及扰动的二范数$\|\rho\|_2$,其中:对抗性指标 $f_{adv}(x^*)$是用来评价生成的对抗样本对路牌识别模型的攻击效果,$f_{adv}(x^*)$越低,攻击效果越好;扰动平滑度 $f_{PSR}(\rho)$是用来评价生成扰动的物理稳定性;扰动的二范数$\|\rho\|_2$ 是用来评价生成扰动的隐蔽性.适应度函数计算公式如下:

$$fit(x^*)=f_{adv}(x^*)+\kappa_1\cdot f_{PSR}(\rho)+\kappa_2\cdot\|\rho\|_2 \tag{6}$$

其中,κ_1,κ_2 是平衡量纲的超参数.κ_1 是为了控制扰动的平滑度,保证物理攻击的有效性.κ_2 是为了控制扰动的隐蔽性.κ_1,κ_2 过大,均容易降低物理攻击成功率;过小时,物理攻击成功率和扰动的隐蔽性效果均会下降.实验发现,κ_1,κ_2 取值范围在 0.001 和 0.8 之间均具有较好的物理攻击效果.本文实验中,分别设为 5×10^{-3} 和 1×10^{-2}.根据适应度函数,我们的寻优目标是搜索到攻击成功率高、扰动的物理稳定性强和扰动的不可见效果好的对抗路牌图像.根据攻击者预设的期望,对抗性指标可分为目标对抗性和无目标对抗性攻击,计算公式如下:

$$f_{adv}(x^*)=\begin{cases}\dfrac{1}{n}\sum_{i=1}^{n}J(f(x_i^*),y_{target}), & \text{目标对抗攻击}\\ \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{score_{\text{true}}}{rank(x_i^*)}\right), & \text{无目标对抗攻击}\end{cases} \tag{7}$$

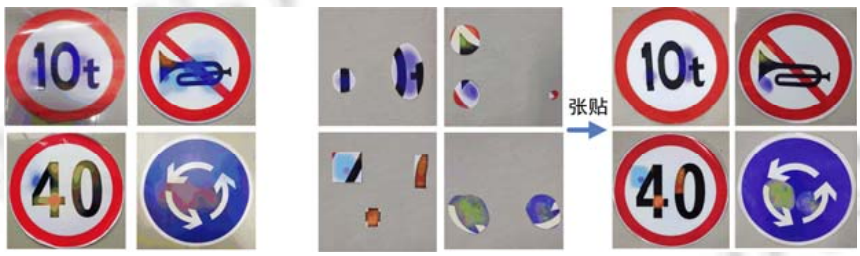
其中,$f(\cdot)$表示路牌分类器的输出,包含所有类标的置信度分数;$J(\cdot)$表示交叉熵函数;y_{target}是目标类类标;n 表示图像变换的类别数,每个粒子通过随机缩放、旋转、亮度变换等操作得到新的 n 张图像,目的是评价扰动的稳定性,在本文实验中,n 的取值为 15;$score_{\text{true}}$是真实类标的置信度分数;$rank(\cdot)$是真实类标的置信度分数排名.

2.5 物理攻击的有效性保证

BPA-PSO 算法致力于实现物理空间中对路牌识别模型的攻击,相比于电子空间的攻击,它的实现更加困难,而一旦实现,危害性也更大.我们采取了以下几种措施来保证物理攻击的有效性.

(1) 使用海报或贴纸

对于添加到路牌上的扰动,其隐蔽性是很重要的,如果太过明显或突出,很容易引起人眼的警觉.本文对于扰动的隐蔽性定义并非指扰动不可见,而是指不引起人们的注意.但是因为物理世界的噪声难以预测,以至于小规模不可见的扰动很容易被破坏掉,无法稳定的存在.因此,采用海报(打印生成的对抗样本,扰动分布于整个路牌图像区域)或贴纸(打印生成的扰动,扰动区域较小)作为扰动的存在形式,可以很好地解决这个问题.如图 3 所示:图 3(a)是打印的海报类型的路牌物理对抗样本,图 3(b)是打印的贴纸类型的路牌物理对抗样本.



(a) 海报类型 (b) 贴纸类型

Fig.3 Physical adversarial examples using posters or stickers

图 3 使用海报或贴纸的物理对抗样本

(2) 增加扰动的平滑性

为了拟合物理世界中物体自然色彩的平滑性,在添加随机扰动和 BPA-PSO 算法对扰动的优化中加入了扰动的平滑性.最后,通过最小化 $f_{PSR}(\rho)$可以使相邻像素点之间的值彼此接近,从而改善对抗样本图像的平滑度.这样不仅减少对抗样本打印时扰动的失真,也使图像传感器能够充分捕捉扰动的特征,促进物理的可实现性.

(3) 增强扰动的存在稳定性

我们充分分析了物理世界环境中的各种因素可能对扰动造成的影响,例如距离、角度、光影变化等,这些都是客观存在的,并且它们的变化可能会导致扰动的失效.因此,BPA-PSO 算法在优化扰动的过程中,通过图像处理中的缩放、旋转、亮度调节来模拟物理真实环境的变化.这样,最后得到的对抗样本就可以在复杂多变的物理环境中仍然具有较强的对抗攻击性.

2.6 算法伪代码

BPA-PSO 算法的伪代码说明如下.

算法伪代码. BPA-PSO 算法.

输入:一张良性路牌图像,适应度阈值 e,粒子数 $numParticles$,最大迭代次数 G_k,当前迭代次数 $iter$;

输出:鲁棒的对抗样本 x^*.

通过在良性路牌图像中添加随机扰动,得到初始对抗样本 $\{x_i'\}$.
初始化:从 $\{x_i'\}$ 中随机选择样本作为初始粒子,第 i 个粒子的历史最优适应值 f_{pi},全局最优适应值 f_g
WHILE $iter \leqslant G_k$ **DO**
  对粒子进行旋转缩放和亮度调节变化;
  根据公式(6)计算粒子的适应度值;
  FOR i=0: *numParticles* **DO**
    IF $fit(x_i^*) < f_{pi}$ **THEN**
      $f_{pi} = fit(x_i^*)$;
      IF $fit(x_i^*) < f_g$ **THEN**
        $f_g = fit(x_i^*)$;
      END
    END
    IF f_g<e **THEN**
      Break;
    END
    根据公式(3)和公式(4)更新粒子速度和位置;
  END
  $iter$=$iter$+1;
END
Return: x^*.

3 实验与分析

3.1 实验平台与评价指标介绍

实验平台环境:i7-7700K 4.20GHzx8(CPU),TITAN Xp 12GiBx2(GPU),16GBx4 内存(DDR4),Ubuntu 16.04(操作系统),Python 3.5,Tensorflow-gpu-1.3(深度学习框架),Tflearn-0.3.2.

为了证明 BPA-PSO 算法的物理可实现性,我们分别在物理世界的实验室模拟场景和真实交通环境中进行了实验验证.通过摄像机采集了大量含有路牌标志的场景图像,并对它们进行路牌标志的检测与识别.同时,为了验证本文算法生成的对抗样本的稳定性,分别对光线、距离、角度等因素进行改变,设计了多种不同的对比场景进行实验.

实验中,以扰动的 2-范数作为电子空间中的扰动评价标准,以攻击成功率作为生成的对抗样本鲁棒性的评价标准,其中,攻击成功率分为电子空间的攻击成功率 ASR_{elec} 和物理空间的攻击成功率 ASR_{phy}:

$$ASR_{elec} = \frac{N_{adv}}{N_{ben}}, ASR_{phy} = \frac{N_{phy}}{N_{adv}} \tag{8}$$

其中,N_{ben} 表示待攻击的良性样本数量,N_{adv} 表示在电子空间中攻击成功的对抗样本数量,N_{phy} 表示在物理空间中攻击成功的对抗样本数量.

3.2 数据集与识别模型介绍

在模型的训练数据中,采用了本团队成员制作的中国路牌数据集(Chinese road sign recognition benchmark,简称 CRSRB),其中包含了 35 类常见的交通路牌标志,共计图片 5 000 张.按照 8:2 的比例划分为训练集和测试集,其中,所有图片大小均为 64×64×3.如图 4 所示为部分数据集展示,采集过程中考虑了不同光照、角度、背景、距离等因素.第 1 行从左到右分别是:限重 10t、禁止鸣笛、限速 40km/h、禁止直行、禁止通行、禁止机动车通

行、连续弯路,第 2 行从左到右分别是:T 字路口、上坡路、步行、环岛行驶、标志指示、注意行人、解除限速 40km/h.

Fig.4 Display of the Chinese road sign

图 4 部分路牌数据集展示

实验选用的路牌识别深度模型结构说明见表 1,包含不同的结构,每种结构分别训练了 5 次,总共得到 15 个识别模型,平均识别准确率大于 95%,与文献[15]中的 95.68%相近,基本满足识别要求.表中“5×5×32”表示卷积核窗口尺寸为 5×5,深度为 32.其中,CHINA-CNN1 采用文献[15]中的结构,使用德国路牌数据集.CHINA-CNN2 和 CHINA-CNN3 在 CHINA-CNN1 的基础上进行了层的修饰,得到结构不同的模型.由于本文使用的路牌识别模型结构与文献[15]提到的 GTSRB-CNN 模型结构相似,因此识别效率在相同的硬件条件下相当.

Table 1 Structure of road sign recognition model

表 1 路牌识别模型的网络结构

Layer type	CHINA-CNN1	CHINA-CNN2	CHINA-CNN3
Conv+ReLu	1×1×3	1×1×3	1×1×3
Conv+ReLu	5×5×32	–	5×5×32
Conv+ReLu	5×5×32	5×5×32	5×5×32
Max Pooling	2×2	2×2	2×2
Dropout	0.75	0.75	0.75
Conv+ReLu	5×5×64	–	5×5×64
Conv+ReLu	5×5×64	5×5×64	5×5×64
Max Pooling	2×2	2×2	2×2
Conv+ReLu	5×5×128	–	5×5×128
Conv+ReLu	5×5×128	5×5×128	5×5×128
Max Pooling	2×2	2×2	2×2
Dropout	0.75	0.75	0.75
Dense(fully connected)+ReLu	1024	1024	–
Dropout	0.75	0.75	–
Dense(fully connected)+ReLu	1024	1024	1024
Dropout	0.75	0.75	0.75
Dense(fully connected)+ReLu	35	35	35
Softmax	35	35	35

3.3 实验结果分析

实验研究了基于 BPA-PSO 算法生成的对抗样本在电子空间、实验室物理空间、户外物理空间(晴天/雨天)等场景下的攻击效果.

(1) 电子空间场景的攻击效果分析

首先评估 BPA-PSO 算法在电子空间中的攻击效果,包括攻击成功率和扰动指标计算.如图 5 所示:图 5(a)所示为原图;图 5(b)所示为 ZOO 算法实现对路牌识别模型的黑盒攻击结果,其中,第 1 列是添加的对抗扰动可视化后的图,第 2 列是添加扰动后的对抗样本路牌图像;图 5(c)所示为 BPA-PSO 攻击方法实现对路牌识别模型的黑盒攻击结果.

进一步,在电子空间中的攻击效果统计如表 2 所示.在基于 ZOO 的物理攻击中,同样考虑扰动平滑操作,以保证物理扰动的可靠性.由实验结果可知:通过 ZOO 算法攻击得到的对抗样本攻击成功率较低;而 BPA-PSO 算法在电子和物理空间的攻击成功率达到了 100%,得到的扰动大小比 ZOO 也更小.这主要是因为 BPA-PSO 算法学习了针对模型的对抗样本的特征分布,在优化过程中保留了对最终要攻击的目标模型具有对抗性的特征.表

2 中物理攻击的对抗路牌图像是在距离小于 5m、倾角小于 5°的情况下拍摄的.

(a) 原图 (b)C&W 电子攻击的扰动及对抗样本 (c)BPA-PSO 电子攻击的扰动及对抗样本

Fig.5 Display of some electronic adversarial road signs image

图 5 部分电子对抗路牌图像展示

Table 2 Attack performance on an electronic space scene

表 2 电子空间场景的黑盒攻击效果

攻击方法	黑盒平均 ASR_{elec}	黑盒平均 ASR_{phy}	平均扰动大小(2-范数)
ZOO	87.30%	53.60%	11.19
BPA-PSO	**100.00%**	**100.00%**	**10.87**

(2) 实验室场景攻击效果分析

在实验室场景中,对路牌识别系统进行了物理攻击,主要采用贴纸和海报的方式来破坏路牌图像的特征,从而使路牌识别系统分类出错.实验测试了 BPA-PSO 算法得到的对抗路牌图像在不同距离和不同倾斜角度情况下的目标/无目标攻击效果.

实验结果见表 3,其中,“5m/0°”表示在距离路牌 5m 处、摄像头旋转 0°的条件下拍摄.

Table 3 Attack performance in indoor scene

表 3 实验室场景下不同距离/倾角的攻击效果

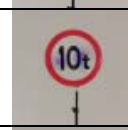
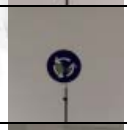
	贴纸				海报	
	无目标攻击		目标攻击		目标攻击	
原始类	禁止鸣笛	环岛行驶	限速 40km/h	限重 10t	限速 40km/h	限重 10t
目标类	~	~	禁止鸣笛	限速 40km/h	禁止鸣笛	限速 40km/h
5m/0°						
7m/0°						
7m/20°						
15m/0°						
15m/20°						
ASR_{phy}	86.70%	93.30%	100.00%	93.30%	100.00%	100.00%

表 3 展示了部分对抗路牌图像,更多对抗样本展示在附录的图 6 中.

Fig.6 Physical adversarial examples in indoor scene

图 6 室内场景下对抗样本拍摄图

根据结果可知,对抗路牌样本在物理攻击测试中具有较高的攻击成功率,说明了物理扰动的有效性和可靠性.对于相同的路牌标识,海报形式的对抗样本的攻击成功率比贴纸形式的更高,这主要是因为海报能够实现比贴纸范围更广的扰动展示.

(3) 真实交通环境(晴天)中的攻击效果分析

在真实交通环境(晴天)中,实验场景设置了包括距离、角度和光影的变化.我们的攻击方式分为有目标攻击和无目标攻击,同时,我们添加扰动的形式包括贴纸和海报.

真实交通环境(晴天)中的无目标攻击实验结果见表 4,在距离/倾角为分别 5m/0°、7m/0°、7m/20°、15m/0°、15m/20°和光影分别为亮、暗的条件下测试对抗样本的物理攻击成功率.

Table 4 Untargeted attack in sunny outdoor scene

表 4 真实交通环境(晴天)中的无目标攻击

添加方式	贴纸							
光影	亮				暗			
原始类	禁止鸣笛	限速 40km/h	限重 10t	环岛行驶	禁止鸣笛	限速 40km/h	限重 10t	环岛行驶
ASR_{phy}	79.0%	83.3%	100%	100%	82.3%	85.0%	100%	100%
添加方式	海报							
光影	亮				暗			
原始类	禁止鸣笛	限速 40km/h	限重 10t	环岛行驶	禁止鸣笛	限速 40km/h	限重 10t	环岛行驶
ASR_{phy}	94.7%	100%	100%	100%	95.6%	100%	100%	100%

更多对抗路牌图像展示在图 7 中.根据实验结果可知,对抗路牌样本在真实的交通环境(晴天)测试中仍然具有较高的攻击成功率,说明 BPA-PSO 算法生成的对抗样本对变化的物理环境具有较强的鲁棒性.同时,海报形式的对抗样本的攻击成功率仍然比贴纸形式的高.

真实交通环境(晴天)中的有目标攻击实验结果见表 5,在距离、角度和光影变化的条件下,测试对抗样本的物理攻击成功率.表中每种情况都拍摄 3 张图片,记录识别结果和对应置信度.

- “tar:0.92”表示被识别为目标类且置信度为 0.92;
- “ori:0.53”表示被识别为原始正确类标且置信度为 0.53;

- “oth:0.33”表示被识别为除原始类和目标类以外的类标且置信度为 0.33.

根据实验结果可知:对抗路牌样本在真实的交通环境(晴天)测试中,尽管环境因素发生变化,大部分的对抗样本仍然能够以较高的置信度欺骗路牌识别系统.说明 BPA-PSO 算法生成的有目标的对抗样本也具有较强的鲁棒性.

Fig.7 Physical adversarial examples in sunny outdoor scene

图 7 室外场景下(晴天)对抗样本拍摄

Table 5 Targeted attack in sunny outdoor scene

表 5 真实交通环境(晴天)中的有目标攻击

添加方式	贴纸				海报	
光影	亮				暗	
对抗路牌图像						
原始类	禁止鸣笛	限速 40km/h	限重 10t	环岛行驶	限速 40km/h	限重 10t
5m/0°	tar:0.92 tar:0.99 tar:0.93	tar:0.98 tar:0.96 tar:0.99	tar:0.99 tar:0.93 tar:0.98	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99
8m/0°	tar:0.84 tar:0.93 tar:0.50	tar:0.93 tar:0.99 tar:0.97	tar:0.77 tar:0.93 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99
8m/20°	tar:0.72 tar:0.97 tar:0.51	tar:0.99 tar:0.98 tar:0.98	tar:0.99 tar:0.93 tar:0.98	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99
15m/0°	tar:0.78 tar:0.51 tar:0.95	tar:0.79 tar:0.57 ori:0.89	tar:0.52 tar:0.84 tar:0.63	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.70 tar:0.98
15m/20°	tar:0.59 tar:0.60 oth:0.33	tar:0.83 ori:0.53 ori:0.50	tar:0.98 oth:0.52 tar:0.88	tar:0.99 tar:0.99 tar:0.99	tar:0.99 tar:0.99 tar:0.99	tar:0.53 tar:0.85 ori:0.54
ASR_{phy}	93.30%	80.00%	93.30%	100.00%	100.00%	93.30%

(4) 真实交通环境(雨天)中的攻击效果分析

在另一真实交通环境中,实验场景设置在了雨天的道路上.由于制作对抗样本的材料是防水的,因此在实验

测试中也获得了不错的攻击效果.

表 6 展示了对抗路牌样本在雨天交通场景下的无目标攻击结果,更多对抗样本展示在图 8 中.从实验结果可以看出:尽管环境因素更加恶劣,对抗样本仍然具有较为可靠的攻击性.

Table 6 Targeted attack in rainy outdoor scene

表 6 真实交通环境(雨天)中的有目标攻击

添加方式	贴纸		海报	
对抗路牌图像				
ASR_{phy}	100.00%	73.20%	100.00%	100.00%

Fig.8 Physical adversarial examples in rainy outdoor scene

图 8 室外场景下(雨天)对抗样本拍摄图

(5) 攻击算法对比

在相同的电子场景和物理场景下,实验还设置了其他 3 种黑盒攻击方法与本文的 BPA-PSO 攻击方法进行对比,分别是 PSO,ZOO[24]和 Boundary[23].实验中,分别使用了每种攻击方法下电子攻击成功率为 100%的对抗样本来做物理场景的攻击.其中,攻击测试的路牌对抗样本包含 5 个类别,共 1 000 张图像.物理场景攻击的路牌图像是在距离小于 5m、倾角小于 5°的情况下拍摄的.

实验结果见表 7,其中,ASR_{elec} 是电子攻击成功率,ASR_{phy} 是物理攻击成功率.

Table 7 Comparison of attack success rate of different attack algorithms

表 7 不同攻击算法的攻击成功率对比

攻击方法	CHINA-CNN1		CHINA-CNN2		CHINA-CNN3	
	ASR_{elec}	ASR_{phy}	ASR_{elec}	ASR_{phy}	ASR_{elec}	ASR_{phy}
BPA-PSO	100%	91.3%	100%	98.3%	100%	95.1%
PSO	100%	5.4%	100%	0%	100%	0%
ZOO	100%	15.7%	100%	4.1%	100%	7.3%
Boundary	100%	7.5%	100%	5.2%	100%	3.8%

从实验结果可以看出:尽管不同攻击方法的对抗样本的电子攻击成功率均为 100%,但在物理场景中,由于物理环境的噪声和物理设备的失真对扰动的破坏,PSO,ZOO 和 Boundary 在物理场景的攻击成功率明显降低;而 BPA-PSO 仍具有较高的物理攻击成功率.可以看出,BPA-PSO 攻击方法生成的对抗样本具有较好的稳定性.

(6) 对抗训练

为了提高路牌识别模型面对对抗攻击的鲁棒性,实验中使用原始良性路牌样本与使用 BPA-PSO 攻击测试集生成的路牌对抗样本混合得到的数据集对路牌识别模型进行对抗训练.混合的数据集中原始良性路牌样本与对抗样本的数量比为 8:2.实验中,对抗训练使用的是 CHINA-CNN3 模型,经过 5 个 epoch 的训练得到鲁棒性较高的路牌识别模型.在计算对抗训练前后的攻击成功率时,我们挑选了 5 类路牌的对抗样本图像共计 1 000 张来测试对抗训练前后的攻击成功率.其中,物理场景攻击测试的路牌图像是在距离小于 5m、倾角小于 5°的情况下拍摄的.

表 8 展示了 BPA-PSO 攻击方法生成的对抗样本和 ZOO 攻击方法生成的对抗样本在路牌识别模型使用 BPA-PSO 生成的对抗样本进行对抗训练前后的攻击效果.

Table 8 Success rate of attack before and after training

表 8 对抗训练前后的攻击成功率

攻击方法	BPA-PSO ASR_{elec}	ZOO ASR_{elec}	BPA-PSO ASR_{phy}
对抗训练前	100%	100%	95.1%
对抗训练后	7.3%	15.7%	4.2%

根据实验结果可知:

- 将对抗样本添加到训练集进行对抗训练后,对抗样本在电子和物理环境的攻击成功率均明显下降;
- 同时,其他攻击方法如 ZOO 的攻击成功率也明显下降.

因此,路牌识别模型的鲁棒性得到了明显的提高.

4 总 结

本文提出了一种基于粒子群优化的路牌识别攻击方法来生成物理可实现的对抗样本,这种攻击方法生成的对抗样本具有较强的鲁棒性和良好的隐蔽性.该攻击方法属于黑盒攻击,可以在不知道模型内部参数的情况下生成对抗样本.

本文分别在实验室场景以及真实道路交通场景下对对抗样本的对抗性进行了检验,由实验结果可见:BPA-PSO 算法生成的对抗样本在复杂多变的物理环境下,能够以高置信度、高欺骗率、高可靠性攻击路牌识别系统.这对于研究如何提高自动驾驶系统中深度模型的鲁棒性具有极大的理论意义与实践价值.

5 未来工作与展望

本文的方法也面临两个挑战.

(1) BPA-PSO 攻击算法需要获得路牌识别模型输出的分类置信度分数,但在某些场合下只能获得输出的类标.所以在今后的工作中,将研究只用模型输出的类标对路牌识别模型进行攻击;

(2) BPA-PSO 算法的扰动计算时间复杂度较大,这是由于通过添加随机扰动来获得粒子群的初始解为后续的寻优过程增加了负担.所以今后的工作中,将改进生成粒子群初始解的方法,优化计算扰动的时间复杂度.

除路牌识别模型外,自动驾驶系统的语音识别、人脸识别、道路安全检测、车辆检测、行人检测等模型都是基于深度学习架构,也面临同样的安全威胁.因此,面对对抗攻击的威胁,设计有效的防御方案是未来研究工作的重点.

目前,在对抗攻击防御方面的主流方法有数据修改、模型修改以及附加网络防御.我们将研究基于 BPA-PSO 算法生成的对抗样本通过对抗训练的防御效果,提高自动驾驶系统中深度学习模型的安全性.

References:

[1] Chen CY, Seff A, Kornhauser A, Xiao JX. DeepDriving: Learning affordance for direct perception in autonomous driving. In: Agarwal S, ed. Proc. of the IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2722−2730. [doi: 10.1109/ICCV.2015.312]

[2] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Chadowitz C, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: CVPR, 2015. 815−823.

[3] Graves A, Jaitly N, Mohamed AR. Hybrid speech recognition with deep bidirectional LSTM. In: Cernocky H, ed. Proc. of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc: IEEE, 2013. 273−278. [doi: 10.1109/ASRU.2013.6707742]

[4] Qing SH. Research progress on Android security. Ruan Jian Xue Bao/Journal of Software, 2016,27(1):45−71 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4914.htm [doi: 10.13328/j.cnki.jos.004914]

[5] Chen JY, Lin X, Wu YY, Chen YX, Zheng HB, Su MM, Yu SQ, Ruan ZY. Double layered recommendation algorithm based on fast density clustering: Case study on Yelp social networks dataset. In: Malavé CO, ed. Proc. of the 2017 Int'l Workshop on Complex Systems and Networks (IWCSN). Doha: IEEE, 2017. 242−252.

[6] Chen JY, Yang DY, Feng ZL. T-cell detector maturation algorithm based on cooperative co-evolution GA. In: Ding YS, ed. Proc. of the 7th Int'l Conf. on Natural Computation. Shanghai: IEEE, 2011. 2295−2299. [doi: 10.1109/ICNC.2011.6022387]

[7] Wang L, Sng D. Deep learning algorithms with applications to video analytics for a smart city: A survey. arXiv Preprint arXiv: 1512.03131, 2015.

[8] Miao QG, Liu RY, Zhao PP, Li YN, Sun EQ. A semi-supervised image classification model based on improved ensemble projection algorithm. IEEE Access, 2018,6:1372−1379.

[9] Liu RY, Song JF, Miao QG, Xu PF, Xue Q. Road centerlines extraction from high resolution images based on an improved directional segmentation and road probability. Neurocomputing, 2016,212:88−95.

[10] Gong MG, Zhou ZQ, Ma JJ. Change detection in synthetic aperture radar images based on deep neural networks. IEEE Trans. on Image, 2012,21(4):2141−2151. [doi: 10.1109/TIP.2011.2170702]

[11] Bao RD, Yu H, Zhu DF, Huang SF, Sun Y, Liu Y. Automatic makeup with region sensitive generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2019,30(4):896−913 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5666.htm [doi: 10.13328/j.cnki.jos.005666]

[12] Wan B, Wang Q, Gao YX. Error diffusion halftone algorithm based on image segmentation. Journal of Xidian University, 2009, 36(3):496−546 (in Chinese with English abstract).

[13] Wang Q, Dong BY, Tian YM. A motion object detection algorithm for MPEG-4 video. Journal of Xidian University, 2007,34(6): 869−872 (in Chinese with English abstract).

[14] Chen JY, Wang Z, Cheng KH, Zheng HB, Pan AT. Out-of-Store object detection based on deep learning. In: Huang L, ed. Proc. of the 2019 11th Int'l Conf. on Machine Learning and Computing. New York: ACM, 2019. 423−428.

[15] Stallkamp J, Schlipsing M, Salmen J, Igel C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks: The Official Journal of the Int'l Neural Network Society, 2012,32:323−332.

[16] Baluja S, Fischer I. Adversarial transformation networks: Learning to generate adversarial examples. arXiv preprint arXiv:1703.09387, 2017.

[17] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

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

[19] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Butler KRB, ed. Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). San Jose: IEEE, 2017. 39−57. [doi: 10.1109/SP.2017.49]

[20] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Bajcsy R, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: CVPR, 2016. 2574−2582.

[21] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Chellappa R, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: CVPR, 2017. 1765−1773.

[22] Su JW, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans. on Evolutionary Computation, 2019, 23(5):828−841. [doi: 10.1109/TEVC.2019.2890858]

[23] Brendel W, Rauber J, Bethge M. Decision-Based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248, 2017.

[24] Chen PY, Zhang H, Sharma Y, Yi JF, Hsieh CJ. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Thuraisingham B, ed. Proc. of the 10th ACM Workshop on Artificial Intelligence and Security (AISec 2017). New York: ACM, 2017. 15−26. [doi: 10.1145/3128572.3140448]

[25] Chen JY, Su MM, Shen SJ, Xiong H, Zheng HB. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. arXiv preprint arXiv:1906.03181, 2019. [doi: 10.1016/j.cose.2019.04.014]

[26] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[27] Sharif M, Bhagavatula S, Bauer L, Reiter MK. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Weippl E, ed. Proc. of the ACM Sigsac Conf. on Computer & Communications Security. New York: ACM, 2016. 1528−1540. [doi: 10.1145/2976749.2978392]

[28] Ma YK, Wu LF, Jian M, Liu FH, Yang Z. Approach to generate adversarial examples for face-spoofing detection. Ruan Jian Xue Bao/Journal of Software, 2019,30(2):469−480 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

[29] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Zeller A, ed. Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&P). Saarbrucken: IEEE, 2016. 372−387. [doi: 10.1109/EuroSP.2016.36]

[30] Cisse M, Adi Y, Neverova N, Keshet J. Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373, 2017.

[31] Dong YP, Liao FZ, Pang TY, Su H, Zhu J, Hu XL, Li JG. Boosting adversarial attacks with momentum. In: Brown M, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 9185−9193.

[32] Sarkar S, Bansal A, Mahbub U, Chellappa R. UPSET and ANGRI: Breaking high performance image classifiers. arXiv preprint arXiv:1707.01159, 2017.

[33] Dorigo M, Stützle T. Ant colony optimization: Overview and recent advances. Handbook of Metaheuristics, 2010,146(5): 227−263.

[34] Kennedy J, Eberhart R. Particle swarm optimization. In: Si J, ed. Proc. of the Int'l Conf. on Neural Networks (ICNN'95). Perth: IEEE, 1995. 1942−1948. [doi: 10.1109/ICNN.1995.488968]

[35] Li XL, Shao ZJ, Qian JX. An optimizing method based on autonomous animats: Fish-swarm algorithm. Systems Engineering—Theory & Practice, 2002,22(11):32−38 (in Chinese with English abstract).

[36] Jiang JG, Zhou JW, Zheng YC, Zhou RS. A double flora bacteria foraging optimization algorithm. Journal of Shenzhen University Science and Engineering, 2014,31(1):43−51.

[37] Karaboga D. Artificial bee colony algorithm. Scholarpedia, 2010,5(3):6915. [doi: 10.4249 / scholarpedia.6915]

[38] Eusuff MM, Lansey KE. Optimization of water distribution network design using the shuffled frog leaping algorithm. Journal of Water Resources Planning & Management, 2003,129(3):210−225. [doi: 10.1061/(ASCE)0733-9496(2003)129:3(210)]

附中文参考文献:

[4] 卿斯汉.Android 安全研究进展.软件学报,2016,27(1):45−71. http://www.jos.org.cn/1000-9825/4914.htm [doi: 10.13328/j.cnki.jos.004914]

[11] 包仁达,庾涵,朱德发,黄少飞,孙瑶,刘偲.基于区域敏感生成对抗网络的自动上妆算法.软件学报,2019,30(4):896−913. http://www.jos.org.cn/1000-9825/5666.htm [doi: 10.13328/j.cnki.jos.005666]

[12] 万波,王泉,高有行.图像分割的误差分散半调算法.西安电子科技大学学报(自然科学版),2009,36(3):496−546.

[13] 王泉,董宝鸳,田玉敏.一种 MPEG-4 视频流的运动目标检测算法.西安电子科技大学学报,2007,34(6):869−872.

[28] 马玉琨,毋立芳,简萌,刘方昊,杨洲.一种面向人脸活体检测的对抗样本生成算法.软件学报,2019,30(2):469−480. http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

[35] 李晓磊,邵之江,钱积新.一种基于动物自治体的寻优模式:鱼群算法.系统工程理论与实践,2002,22(11):32−38.

陈晋音(1982－),女,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为人工智能安全,深度学习,图数据挖掘,进化计算.

陈治清(1998－),男,硕士生,主要研究领域为深度学习,数据挖掘.

郑海斌(1995－),男,博士生,CCF 学生会员,主要研究领域为人工智能安全,深度学习应用,数字图像处理.

沈诗婧(1996－),女,硕士生,主要研究领域为计算机视觉,人工智能.

苏蒙蒙(1994－),女,硕士生,主要研究领域为深度学习,人工智能安全.