文章编号:1671-4598(2022)07-0255-06 DOI:10.16526/j.cnki.11-4762/tp.2022.07.038 中图分类号:TP18 文献标识码:A

基于决策的目标检测器黑盒对抗攻击方法

付 平,郭 玲,刘 冰,朱玉晴,凤 雷
(哈尔滨工业大学 电子与信息工程学院,哈尔滨 150001)

摘要:深度神经网络在目标检测领域有大量的应用已经落地,然而由于深度神经网络本身存在不可解释性等技术上的不足,导致其容易受到外界的干扰而失效,充分研究对抗攻击方法有助于挖掘深度神经网络易失效的原因以提升其鲁棒性;目前大多数对抗攻击方法都需要使用模型的梯度信息或模型输出的置信度信息,而工业界应用的目标检测器通常不会完全公开其内部信息和置信度信息,导致现有的白盒攻击方法不再适用;为了提升工业目标检测器的鲁棒性,提出一种基于决策的目标检测器黑盒对抗攻击方法,其特点是不需要使用模型的梯度信息和置信度信息,仅利用目标检测器输出的检测框位置信息,策略是从使目标检测器定位错误的角度进行攻击,通过沿着对抗边界进行迭代搜索的方法寻找最优对抗样本从而实现高效的攻击;实验结果表明所提出的方法使典型目标检测器 Faster R-CNN 在 VOC2012 数据集上的 mAR 从 0.636 降低到 0.131,mAP 从 0.801 降低到 0.071,有效降低了目标检测器的检测能力,成功实现了针对目标检测器的黑盒攻击。

关键词:深度神经网络;目标检测器;黑盒攻击;对抗样本;对抗边界

Decision-based Black Box Adversarial Attack Method for Target Detector

FU Ping, GUO Ling, LIU Bing, ZHU Yuqing, FENG Lei
(School of Electronic and Information Engineering, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Deep neural network has been widely applied in the field of object detection. However, due to the poor interpretability and other technical deficiencies of deep neural network, it is easy to be invalidated by external interference. Full research on adversarial attack methods is helpful to explore the reason for the invalidation of deep neural network and improve its robustness. At present, most of the adversarial attack methods need to use the gradient information of the model or the confidence information of the model output, but the object detectors used in the industry usually do not fully open their internal information and confidence information, which leads to that the existing white box attack methods are no longer applicable. To enhance the robustness of industrial object detector, a decision-based black box adversarial attack method for object detector is proposed. The characteristics of this method does not need to use the gradient information and confidence information of the model, only use the object detector output detection box position information. The strategy of this method is to make the object detector locate wrong and attack it, and to find the optimal adversarial examples by iterative search along the adversarial boundary so as to achieve efficient attack. Experimental results show that the proposed method reduces mAR from 0.636 to 0.131 and mAP from 0.801 to 0.071 on VOC2012 data set of typical object detector Faster R-CNN, which effectively reduces the detection ability of object detector and successfully achieves the black box attack on the object detector.

Keywords: Deep neural network; object detector; black-box adversarial attack; adversarial example; adversarial boundary

0 引言

随着深度学习理论知识的不断扩大及硬件计算资源的发展,深度神经网络受到了广泛的研究和应用。深度神经网络快速发展使得其在图像分类[1]、目标检测[2]、图像分割[3]和目标跟踪[4]等各项任务中均有出色的表现。在目标识别领域中,深度神经网络有大量的应用已经落地,比如日常生活中的智能手机利用人脸识别技术[5]进行屏幕解锁。深度神经网络已经渗透进我们的日常生活中,而其本身具有的不可解释性使得其行为难以解释和控制[6],因此关于其鲁棒性和稳定性的研究大量涌现。一些研究发现尽管深度神经网络具有准确性高和应用广泛等优势,但其在受到一些外界干扰时易失效。

Szegedy 等[6]首次发现在输入图像上添加一些特定的人眼无法察觉的扰动可以使深度神经网络模型无法正确识别,这种特定的微小扰动称为对抗扰动,添加对抗扰动后的图像称为对抗样本,生成对抗样本并使模型识别错误的算法称为对抗攻击算法。对抗攻击现象可能会对我们的人身及财产安全造成巨大的影响,只有充分研究对抗攻击算法,

收稿日期:2022-05-03; **修回日期**:2022-05-19。
基金项目:国家自然科学基金(62171156)。
作者简介:付 平(1965-),男,黑龙江人,博士,教授,主要从事自动测试系统方向的研究。
刘 冰(1982-),男,黑龙江人,博士,副教授,主要从事自动测试系统、图像处理处理和 FPGA 加速方向的研究。
通讯作者:凤 雷(1978-),男,黑龙江人,博士,副教授,主要从事自动测试系统研究。
引用格式:付 平,郭 玲,刘 冰,等.基于决策的目标检测器黑盒对抗攻击方法[J].计算机测量与控制,2022,30(7):255-260.

才能发现深度神经网络易失效的原因，促进对抗攻击防御方法的完善和深度神经网络安全性方面研究的发展，从而有针对性的构建更加鲁棒的深度神经网络模型。目标检测作为计算机视觉的一个热门方向，有着越来越多的应用，现有的针对目标检测器的对抗攻击方法都是基于梯度的攻击或是基于置信度的攻击，而工业界应用的目标检测通常只为用户提供最终的决策信息，不会提供过多模型内部信息和带有置信度的识别结果，所以以上攻击方法将不再适用。

为了得到更适用于工业界应用的目标检测器的对抗攻击方法，本文提出了一种基于决策的目标检测器黑盒对抗攻击方法，该方法通过使目标检测器定位错误来进行攻击，在迭代优化时，沿着对抗样本与非对抗样本的边界执行游走，保证其定位错误的前提下，减小对抗样本的扰动。本文提出的对抗攻击方法不需要用到目标检测器的梯度信息以及输出标签和检测框的置信度信息，仅利用其输出的检测框的位置信息即可实现高效的攻击。

1 相关工作

近年来，研究人员在对抗攻击方向进行了大量的研究，提出了多种对抗攻击算法。

Szegedy 等人[6]在 2014 年首次在图像分类领域发现了对抗攻击现象，并提出 L-BFGS 方法来生成对抗样本。研究者在 L-BFGS 基础上又相继提出效果更好的 FGSM 方法[7]，DeepFool 方法[8]、C&W 方法[9]和 JSMA 算法[10]，以上方法均为白盒攻击方法，需要知道模型的内部结构，利用其损失函数的梯度信息才能实现攻击。Chen 等人在 2017 年提出了一种黑盒攻击方法——ZOO（Zeroth Order Optimization）方法[11]，这种方法不需要知道模型的梯度信息，而是通过与模型多次进行交互，利用大量模型输出的置信度信息来对受攻击模型进行梯度估计，从而生成对抗样本。Su 等提出了 ONE-PIXEL 黑盒攻击方法[12]，仅通过改变图像的一个像素就能实现对模型的攻击，该方法应用微分进化来找到最优解。以上黑盒攻击方法均需利用模型输出的置信度信息，而大多工业界应用的模型只会为用户提供最终的决策，Brendel 等提出了一种边界攻击方法[13]，能够高效攻击更接近现实场景中的黑盒模型，该方法不需要使用梯度或置信度信息，利用决策边界的几何属性，在获得极少信息量的情况下对模型进行高效的攻击。

对抗攻击方法在图像分类领域取得了一定的研究成果后，研究人员开始将对抗攻击方法扩展到目标检测领域。2017 年，Lu 等人[14]提出了 DFool 算法，是一种针对 Faster R-CNN[15]的白盒对抗攻击方法，其利用 Faster R-CNN 在所有 STOP 交通标志上进行得分测试，通过最小化所有图像的平均预测得分，设计出 STOP 交通标志的对抗样本。Xie 等人[16]针对目标检测器的分类损失函数提出了 DAG 攻击方法，该方法首先得到每个目标物体的正确类别，然后再为目标物体设置一个不正确的标签，通过迭代增大不正确的标签的置信度，同时减小正确标签的置信度，最终使得检测器对输入图片的所有感兴趣区域（RoI，region of interest）都分类错误。Wang 等人[17]提出了一种针对 YOLO 目标检测器[18]的对抗攻击方法——Daedalus，通过攻击 YOLO 的非极大值抑制机制来使 YOLO 检测错误。Wu 等人在 2019 年提出了一种针对目标检测器的黑盒攻击方法——G-UAP[19]，其是在 UAP 方法[20]的基础上进行了改进，通过诱导 RPN 网络[15]将前景目标误认为是背景，即降低图片中前景目标的置信度，同时增加背景的置信度从而实现攻击。2020 年，Chow 等人[21]从目标检测器的输出结构考虑，提出了 3 种有针对性的对抗攻击方法，称为 TOG，其利用模型的损失函数，通过 3 种类型的有目标攻击来欺骗目标检测器，包括目标消失、目标制造和目标标签错误。

仅利用模型的决策信息进行对抗攻击的方法目前只在图像分类领域有相关研究，现有的针对目标检测器的对抗攻击方法都需要知道模型的内部结构或者知道其输出的置信度信息才能实现高效的攻击，而工业界应用的目标检测器通常不会为用户提供过多的信息，因此本文提出了一种适用于工业界应用的目标检测器的对抗攻击方法。

2 对抗边界与对抗样本标准的设计

本文提出的对抗攻击方法通过沿着对抗边界游走以找到最优对抗样本，使目标检测器定位准确度降低。为了更好的描述本文方法，现给出定位准确度评价指标以及对抗边界与对抗样本标准的设计。

2.1 定位准确度评价指标

目标检测器的主要任务不仅是分类，同时还需要定位，目标检测器对于单个目标的定位好坏通常用 *IoU* 来评价。*IoU* 用于衡量预测框和人工标注的真实框的重合程度，其计算方法如下：

$$IoU = \frac{\text{预测回归框} \cap \text{真实回归框}}{\text{预测回归框} \cup \text{真实回归框}} \tag{1}$$

即预测回归框与真实回归框的交集比上预测回归框与真实回归框的并集。

通常，我们会设置一个 *IoU* 门限值，大于这个门限值的检测框我们就认为其检测到了目标，而小于这个门限值我们就认为其没有检测到目标，当 *IoU* 门限值设置为 0.5 时的检测效果如图 1 所示。

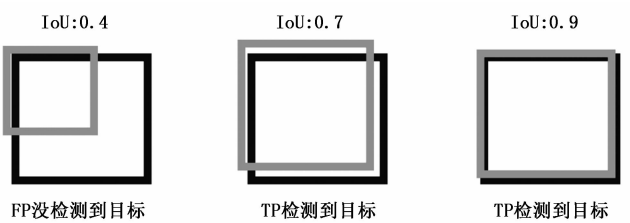


图 1 *IoU* 门限值为 0.5 时检测效果

利用对抗攻击方法生成对抗样本是针对整张图片而言的，而不是一个目标，只能评价单个目标的定位好坏，对于一张图片所有目标的定位结果我们引入平均作为评价指

标，其计算公式如下：

$$avg\_loU = \frac{IoU1 + IoU2 + \cdots + IoUN}{N} \tag{2}$$

其中：$IoU1$、…、$IoUi$、…、$IoUN$ 分别为第 i 个真实框所匹配的最大 IoU 值。

以图 2 为例，两个黑色的边界框为我们标注的真实边界框，4 个灰色的边界框为模型预测的边界框，我们的目标就是为找到分别与所有真实边界框 IoU 最大的预测框，并记录 IoU 值，然后计算平均值。以图 2 为例，我们首先选取真实框 1，并遍历 4 个预测框，找到与真实框 1 的 IoU 最大的预测框为预测框 1，记录 IoU 此为 $IoU1$；然后选取真实框 2，由于预测框 1 已经被匹配了，我们就不再匹配它，所以遍历剩下的 3 个预测框，找到与真实框 2 的 IoU 最大的预测框，记录 IoU 此为 $IoU2$（此时没有预测框与真实框 2 相交，所以 $IoU2$ 为 0）。所以平均 IoU 为（$IoU1 + IoU2$）/2。

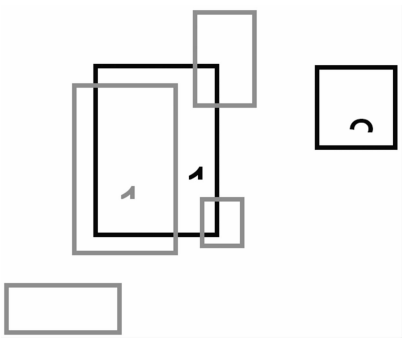


图 2　计算 IoU 示例

2.2　对抗边界

在一个目标检测神经网络中，以图像 A 的人工标注框作为真实框，所有平均 IoU 小于 50%的图像所在的区域，就为图像 A 的对抗区域 G_A；反之，为非对抗区域 $\overline{G_A}$，公式如下：

$$\begin{aligned} G_A &= \{P_i \mid avg\_IoU_A(P_i) > 50\%\} \\ \overline{G_A} &= \{P_i \mid avg\_IoU_A(P_i) < 50\%\} \end{aligned} \tag{3}$$

在一个目标检测神经网络中，图像 A 的对抗边界指图像 A 的对抗区域中最靠近其非对抗区域的图像的集合，即图像 A 的对抗区域的最内层，进行极其微小的移动都会进入到非对抗区域。图像 A 的对抗边界为：

$$\begin{cases} B_A = \{P_i \mid avg\_IoU_A(P_i) > 50\%\} \wedge \\ avg\_IoU_A(P_i + \varepsilon) < 50\% \} \\ \varepsilon \to 0 \end{cases} \tag{4}$$

2.3　对抗样本标准

在本文提出的攻击方法中，判断图像为对抗样本的标准为图像的平均 IoU 小于 50%。当受攻击的目标图像的平均 IoU 从 50%以上降低到 50%以下时，我们就认为目标检测器定位失效了，其对该图片的定位发生了错误。

3　目标检测器决策攻击方法

充分研究对抗攻击算法有助于对对抗攻击防御方法进行完善，增强深度神经网络模型的鲁棒性，为了挖掘工业应用的目标检测器易失效的原因以提高其鲁棒性，本文提出了一种基于决策的目标检测器黑盒对抗攻击方，通过使目标检测器定位错误来实现针对目标检测器的对抗攻击，借鉴了边界攻击[13]的思想，利用迭代搜索的方法寻找最优对抗样本。

3.1　总体流程

本文首次将边界攻击的思想应用于目标检测任务，提出了针对目标检测器的决策攻击方法。该方法总体流程图如图 3 所示，在得到原始输入图像后，本文提出的方法需要使用已经对抗的样本进行初始化，首先粗略地寻找无扰动限制的初始对抗样本，为了减小扰动大小，将初始对抗样本朝着原图的方向进行移动直到到达对抗边界附近并保证图像仍为对抗样本，然后将当前对抗样本沿着对抗边界执行迭代游走，游走的方向需要为与原始图像距离更近的方向，同时仍停留对抗区域中，从而保证样本仍是对抗样本的同时降低与原始图像的距离，为了使避免扰动难以收敛的问题，每迭代固定次数后将进行超参数调整，当迭代次数达到我们所设置的次数上限后，得到最终对抗样本。其中，本文采用欧氏距离衡量两个图像之间的距离，假设图像 A 的像素点矩阵为 $\boldsymbol{X}$，每一个像素点值分别为 x_1，x_2，…，x_n，图像 B 的像素点矩阵为 $\boldsymbol{Y}$，每一个像素点值分别为 y_1，y_2，…，y_n，则图像 A 和图像 B 之间的距离公式如下：

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

3.2　攻击对象与初始对抗样本生成

本文提出的攻击方法主要攻击的对象是平均 IoU 大于 50%的图像，即原始输入图像的平均 IoU 大于 50%。通常认为目标检测模型对平均 IoU 小于 50%的图像本身定位准确度就较低，所以没有必要进行攻击。

本文方法需要使用已经对抗的样本进行初始化，在寻找初始对抗样本的过程中，我们首先需要在测试集中找到检测结果与输入图像的真实回归框平均 IoU 小于 0.5 的图像，作为干扰图像，与原始图像进行叠加。令 α 为叠加系

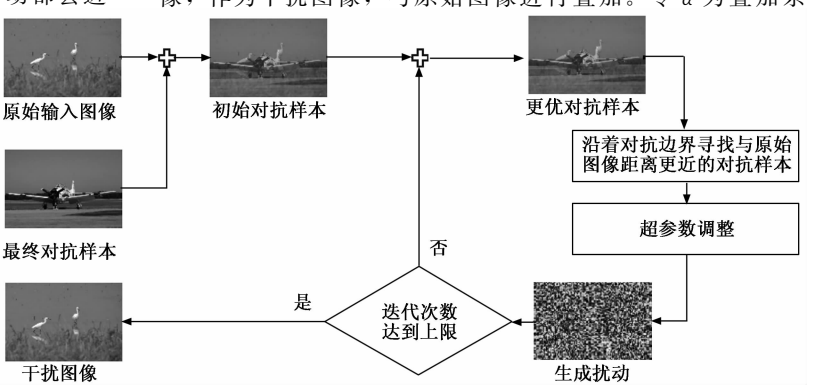


图 3　针对目标检测器的决策攻击方法流程图

数，初始值为 0.5，通过多次二分查找，保证仍是对抗样本的同时尽可能降低 α 的值，以使初始对抗样本与原始图像的距离尽可能的小一些，减少后续迭代次数，搜索到合适的叠加系数 α 后，生成初始对抗样本，其公式如下：

$$\hat{x}^0 = a \cdot x_{\text{reversed}} + (1-\alpha) \cdot x \tag{6}$$

其中：$\hat{x}^0$ 为初始对抗样本，x_{reversed} 为干扰图像，x 为原始输入图像。

3.3 对抗样本搜索策略

本文提出的方法通过沿着对抗边界多次迭代搜索，使对抗样本和原始图像逐渐接近，以找到更优的对抗样本。为了保证每一次迭代都能得到更优的对抗样本，在第次迭代中，产生的扰动 λ^k 需受到以下约束：

1）保证新样本合法：

$$(\hat{x}^{k-1} + \lambda^k) \in [0,255] \tag{7}$$

2）新样本仍为对抗样本：

$$avg\_IoU(\hat{x}^{k-1} + \lambda^k) < 50\% \tag{8}$$

3）新样本与原始图像的距离减小：

$$D(x,\hat{x}^{k-1} + \lambda^k) < D(x,\hat{x}^{k-1}) \tag{9}$$

其中：x 为原始输入图像，$\hat{x}^{k-1}$ 为第 $k-1$ 次迭代得到的对抗样本。

我们将一次迭代过程分为两步移动，如图 4 所示，第一步进行正交移动，即沿着以原始图像为中心的超球面上先走一步，这一步保证需要保证距离原始图像的距离保持不变；第二步在第一步的基础上朝着原始图像点的方向前进，这一步的主要作用就是使对抗样本和原始图像之间的距离减小。我们将一次迭代过程分为两步移动的主要原因是 IoU 大于 50%的区域并不是以原始样本为中心的一个圆，这意味着即使我们走到了区域边界也不一定是距离原始图像最近的地方，甚至扰动可能还非常大，如图 4 所示的对抗样本点，若我们直接朝着原始图像点的方向进行移动，那么能移动的距离非常有限，稍微增加步长可能就会到 IoU 大于 50%的区域，此时，不管迭代多少次，对抗样本与原始样本的距离都无法再减小了。而如箭头所示，如果我们按照我们约束的方向分为两步走，我们将会到达离原始图像更近的地方，同时使样本仍停留在对抗区域。

若我们在第 k 次迭代时失败了，即走进了非对抗区域，那么我们这一次迭代就不进行移动，对抗样本依然保持我们上一次（第 $k-1$ 次）迭代后的状态，调整方向或步长，然后继续下一次迭代。

3.4 超参数更新方式

针对目标检测器的决策攻击方法在迭代过程中主要对两个超参数进行动态调整，一个是正交移动的步长 δ，另一个是朝原始图像前进的步长 η。为了使算法更快的收敛，我们在迭代的过程中自动调整超参数，调整策略如下：在 n 次迭代中，记录正交移动成功的次数 α，朝原始图像移动成功的次数 β，α/n 为正交移动的成功率，其值越大，说明正交方向距离对抗边界越远，为了提高算法收敛速度，需增大 δ，反之，α/n 的值过小，说明多次移动都越过了决策边界进入了非对抗区域，则需减小 δ；朝原始图像移动是在正交移动的基础上进行的，所以其成功率需要在正交移动成功的条件下进行计算，即 β/α，其值越大，说明朝原始图像的方向距离对抗边界越远，需增大 η 以提高算法迭代效率，相反，当 β/α 的值越小，说明多次移动都失败了，需减小 η。

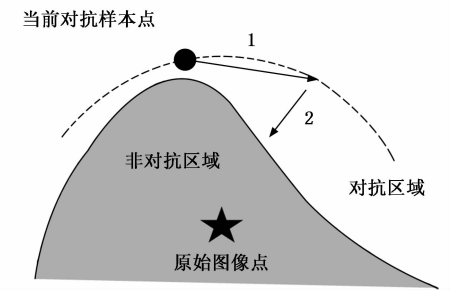


图 4 一次迭代移动过程

为了确定步长缩放系数，对其进行了对比实验，实验以样本收敛所需的最少迭代次数为标准来评价算法的收敛速度，从而确定最佳参数设置。实验结果显示，步长缩放系数为 3 时，算法的收敛速度最快，同时，移动成功率在 0.2～0.6 时算法具有相对稳定的收敛速度，因此，步长调整时机（即成功率为多少时进行调节）设置为 0.2 和 0.6。动态调整步长的具体策略如下：每迭代 25 次自动调整一次步长，若正交移动成功率大于 0.6，那么我们就将 δ 增大为 3δ，若我们的正交移动成功率小于 0.2，就将 δ 减小为 $\delta/3$，若成功率在 0.2 和 0.6 之间，那么我们就维持该步长不变；同理，朝原图移动的步长也是同样的调整策略，如果成功率大于 0.6，就增大 η 到原来的 3 倍，如果成功率小于 0.2，就减小 η 为原来的 1/3 倍，若成功率在 0.2 和 0.6 之间，就维持 η 不变，然后进入下一个 25 次的迭代。

4 实验结果与分析

4.1 实验设置

1）数据集。

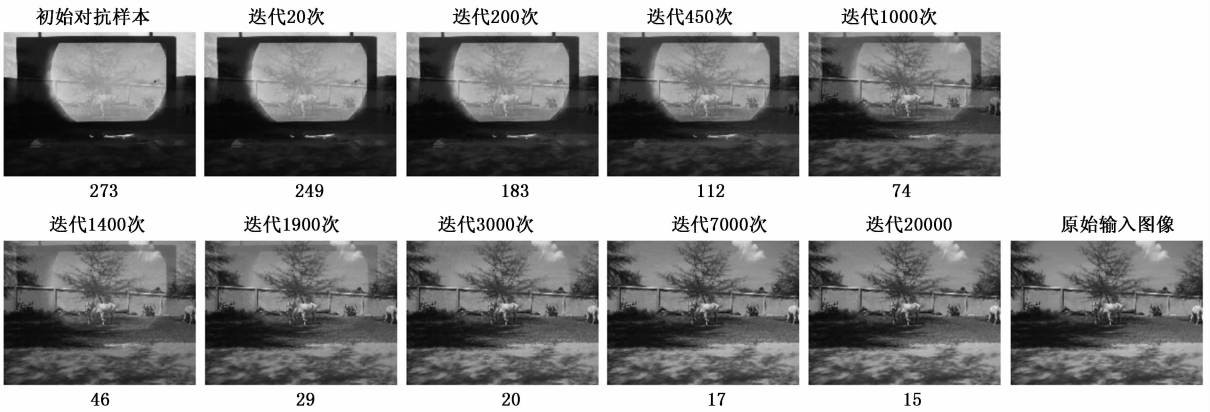
PASCAL VOC2012 数据集是 PASCAL VOC 大赛中所提供的提供了一整套标准化的优秀的数据集，是目标检测任务的基准数据之一，在目标检测任务中被频繁使用。VOC2012 数据集总共分为 20 类，共 11 540 张图片，其中训练集有 5 717 张图片，验证集有 5 823 张图片。本文利用训练集中所有图片训练目标检测模型，并从验证集中随机抽取 256 张图片作为原始图片，利用本文提出的方法生成对抗样本，并测试 mAR 与 mAP。

2）目标检测模型。

本文使用在 VOC2012 数据集上训练的 Faster-Rcnn 目标检测模型作为受攻击模型，主干网络（backbone）部分采用 ResNet50 和 FPN 组合的网络结构。模型的阈值设置为 0.5，即置信度高于 0.5 时才认为检测到了目标。

4.2 实验结果及分析

本实验在 VOC2012 验证集中随机抽取的 256 张图片进



注：图像下方数字表示对抗样本与原始图像之间的距离

图 5　本文方法生成对抗样本的具体过程

行了对抗样本生成，图 5 展示了利用本文方法生成对抗样本的具体过程，从图中可以看出，随着迭代次数的增加，对抗样本与原始图像之间的距离逐渐减小。在迭代初期，距离减小的速度较快，到迭代后期时，对抗样本逐渐趋于收敛，距离减小的速度变缓，迭代 20 000 次时，对抗样本已经收敛。

由于目标检测问题中的每个图像都可能具有不同类别的不同目标，模型的分类和定位都需要进行评估，因此，在图像分类问题中所使用的标准度量不能直接应用于目标检测问题。平均召回率（AR，average recall）表示某一类别检出正样本占实际正样本总数的比例，即查全率，平均精度（AP，average precision）表示某一类别被分为正样本的样本中实际为正样本的比例，即查准率。为了验证本文方法的有效性，实验采用全类平均召回率（mAR，mean average recall）和全类平均精度（mAP，mean average precision）评估目标检测模型的检测准确性，即对所有类别计算 AR 与 AP 的平均值，在 VOC2012 数据集上测得的 mAR 与 mAP 的值如表 1 所示。从表 1 可以看出，本文提出的攻击方法使目标检测器的 mAR 从 0.636 降低到 0.131，mAP 从 0.801 降低到 0.071，表明了模型在受到本文提出的攻击方法攻击后，其检测能力被有效降低，该方法使得目标检测模型失效，且具有较好的攻击性能。

表 1　实验结果对比

	mAR	mAP
干净样本	0.636	0.801
本方法	0.131	0.071

本文提出的攻击算法在 VOC2012 数据集取得了理想的攻击效果。为了验证生成的对抗样本只会使目标检测器误判而不会影响“人眼”的判断，本实验进行了定性分析，图 6 展示了本文方法生成的对抗样本示例，左边是模型对原始图像的检测结果，右边是模型对对抗样本的检测结果。模型受到攻击后，已经无法检测出第一行样本中的显示器，第二行样本的两个牛的目标仅检测出了一个，第三行样本的中的人也没有检测到，第四行样本中的两只羊仅检测出了一只。可以看到对抗样本在视觉上与原始图像十分接近，人眼无法察觉出干扰的存在，因此对抗样本不会使人判断错误，但使得模型对其检测结果的平均 *IoU* 均降低到了 50%以下，造成了模型的定位准确度降低。

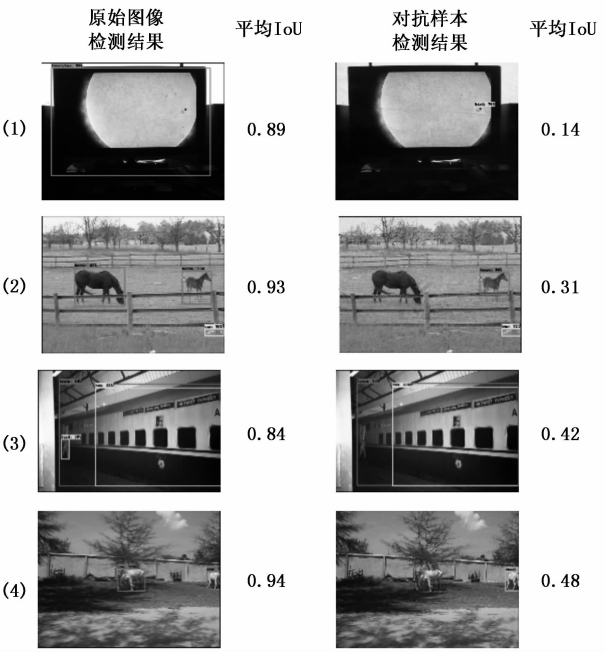


图 6　生成对抗样本示例

4.3　超参数选择

为了确定步长缩放系数对实验的影响，在 VOC 验证集中随机选取了三张图片作为测试样本，利用本文提出的攻击方法进行攻击，除了步长缩放系数外，其他参数均相同设置，通过调整步长缩放系数，进行对比实验以观察其对对抗样本收敛所需迭代次数的影响，实验结果如图 7 所示。由曲线图可以发现，随着步长缩放系数的增大，三张测试集所生成的对抗样本收敛所需的迭代次数均先减小后又增大，步长缩放系数设置为 3 左右时，三张对抗样本所需迭代次数均最少。因此步长缩放系数选取为 3，可以有效提高

对抗样本的生成效率。

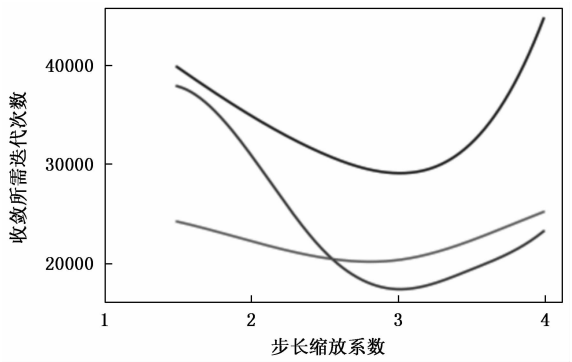


图 7 收敛所需的迭代次数与步长缩放系数关系图

5 结束语

为了揭示真实世界中大多只提供最终决策的商业目标检测器的弱点以提高其鲁棒性，本文提出了一种针对目标检测器的决策攻击方法，该方法的特点是不需要知道模型的梯度信息和置信度信息，仅需利用目标检测器输出的检测框的位置信息，从使目标检测器定位错误的角度实现高效的攻击。但由于该方法了解模型的信息有限，需通过大量的访问模型对对抗样本进行迭代从而生成最终对抗样本，该过程需耗费大量的时间，对抗样本的生成速度还需进一步提升。

参考文献：

[1] DENG J, DONG W, SOCHER R, et al. Imagenet: A largescale hierarchical image database [C] // IEEE conference on computer vision and pattern recognition, 2009: 248－255.

[2] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580－587.

[3] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C] //Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 3431－3440.

[4] DANELLJAN M, HAGER G, SHAHBAZ KHAN F, et al. Convolutional features for correlation filter based visual tracking [C] //Proceedings of the IEEE international conference on computer vision workshops, 2015: 58－66.

[5] TAIGMAN Y, YANG M, RANZATO M A, et al. Deepface: Closing the gap to human-level performance in face verification [C] // Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 1701－1708.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C] // International conference on learning representations, 2014.

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversaryal examples [C] // International conference on learning representations, 2015.

[8] 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, 2016: 2574－2582.

[9] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C] //IEEE symposium on security and privacy, 2017: 39－57.

[10] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in advers-arial settings [C] //IEEE europeanan sympos-ium on security and privacy, 2016: 372－387.

[11] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models [C] //Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017: 15－26.

[12] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE transactions on evolutionary computation, 2019, 23 (5): 828－841.

[13] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning Models [C] // International conference on learning representations, 2018.

[14] LU J, SIBAI H, FABRY E. Adversarial examples that fool detectors [J]. arXiv preprint, 2017: 1712.02494.

[15] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-Time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39 (6): 1137－1149.

[16] XIE C, WANG J, ZHANG Z, et al. Adversarial examples for semantic segmentation and object detection [C] //Proceedings of the IEEE international conference on computer vision, 2017: 1369－1378.

[17] WANG D, LI C, WEN S, et al. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples [J]. IEEE transactions on cybernetics, 2021: 1－14.

[18] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, realtime object detection [C] //Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779－788.

[19] WU X, HUANG L, GAO C. G-uap: Generic universal adversarial perturbation that f-oolsrpn-based detectors [C] //Asian conference on machine learning, 2019: 1204－1217.

[20] 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.

[21] CHOW K H, LIU L, LOPER M, et al. Adversarialobjectness gradient attacks in real-time object detection systems [C] // 2020 second IEEE international conference on trust, privacy and security in intelligent systems and applications, 2020: 263－272.