一种基于局部扰动的图像对抗样本生成方法

王辛晨, 苏秋旸, 杨邓奇, 陈本辉, 李晓伟

大理大学 数学与计算机学院 大理 中国 671000

摘要 近年来, 随着人工智能的研究和发展, 深度学习被广泛应用。深度学习在自然语言处理、计算机视觉等多个领域表现出良好的效果。特别是计算机视觉方面, 在图像识别和图像分类中, 深度学习具备非常高的准确性。然而越来越多的研究表明, 深度神经网络存在着安全隐患, 其中就包括对抗样本攻击。对抗样本是一种人为加入特定扰动的数据样本, 这种特殊样本在传递给已训练好的模型时, 神经网络模型会输出与预期结果不同的结果。在安全性要求较高的场景下, 对抗样本显然会对采用深度神经网络的应用产生威胁。目前国内外对于对抗样本的研究主要集中在图片领域, 图像对抗样本就是在图片中加入特殊信息的图片数据, 使基于神经网络的图像分类模型做出错误的分类。已有的图像对抗样本方法主要采用全局扰动方法, 即将这些扰动信息添加在整张图片上。相比于全局扰动, 局部扰动将生成的扰动信息添加到图片的非重点区域, 从而使得对抗样本隐蔽性更强, 更难被人眼发现。本文提出了一种生成局部扰动的图像对抗样本方法。该方法首先使用 Yolo 目标检测方法识别出图片中的重点位置区域, 然后以 MIFGSM 方法为基础, 结合 Curls 方法中提到的先梯度下降再梯度上升的思想, 在非重点区域添加扰动信息, 从而生成局部扰动的对抗样本。实验结果表明, 在对抗扰动区域减小的情况下可以实现与全局扰动相同的攻击成功率。

关键词 对抗样本; 局部扰动; 目标检测; 神经网络

中图法分类号 TP309.1 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.11.06

A Method of Image Adversarial Sample Based on Local Disturbance

WANG Xinchen, SU Qiuyang, YANG Dengqi, CHEN Benhui, LI Xiaowei

School of Mathematics and Computer, Dali University, Dali 671000, China

Abstract In recent years, with the research and development of artificial intelligence, deep learning has been widely used. Deep learning has shown good results in many fields such as natural language processing and computer vision. Especially in computer vision, deep learning has a very high accuracy in image recognition and image classification. However, more and more studies show that deep neural networks have security risks, including adversarial sample attack. Adversarial sample is a kind of data sample artificially added with specific perturbations. When this special sample is passed to the trained model, the neural network model will output different results from the expected results. In scenarios with high security requirements, adversarial samples will obviously pose a threat to applications using deep neural networks. At present, the research on adversarial samples at home and abroad is mainly focused on the field of images. Image adversarial samples are the image data with special information added to the pictures, so that the image classification model based on neural networks can make the wrong classification. The existing image adversarial sample methods mainly use the global disturbance method, that is, the disturbance information is added to the whole image. Compared with the global disturbance, the local disturbance adds the generated disturbance information to the non-key area of the image, which makes the adversarial sample more hidden and harder to be found by the human eye. In this paper, an image adversarial sample method for generating local disturbances is proposed. The method first used Yolo object detection method to identify the focal areas in images. Then, based on the MIFGSM method and the thought of gradient descent followed by gradient ascending mentioned in the Curls methods, the perturbation information was added to the non-focal areas to generate local perturbation adversarial samples. The experimental results show that the same attack success rate as the global disturbance can be achieved when the anti-disturbance area is reduced.

Key words adversarial sample; local disturbance; target detection; the neural network key word

通讯作者: 李晓伟, 博士, 副教授, 硕士生导师, Email:lixiaowei_xidian@163.com。

本课题得到国家自然科学基金(No. 61902049, No. 62262001)资助

收稿日期: 2022-07-07; 修改日期: 2022-10-13; 定稿日期: 2022-10-13

1 引言

随着人工智能的高速发展, 深度学习被广泛的应用到了自然语言处理、语音识别、计算机视觉等多个领域并展现出良好的效果。特别是计算机视觉方面, 在图像识别和图像分类中, 深度学习具备非常高的准确性, 甚至在一些特殊情况下表现出超出人类的工作能力。因此越来越多的研究人员将深度神经网络应用到各个生产生活领域。由于深度神经网络对于数据的分析理解方式与人类存在明显差异, 模型的神经元数量大, 参数多, 且层次之间连接方式众多, 许多时候深度学习的工作原理和工作方式仍然缺乏可解释性[1]。因为深度神经网络结构复杂, 其最终的结果有时候会与预期的结果存在差距, 甚至远超人们的预想。这就导致了一些“奇怪的”事情发生。2014 年在 Szegedy 等人的论文中提到, 对于一些加入特殊扰动的图像, 深度神经网络表现出很高的脆弱性, Szegedy 将这样的图像称作“对抗样本”。对于图像来说, 这样的扰动通常很小, 难以察觉, 但它们完全欺骗了深度学习模型[2]。如图 1 所示, 将没被加入扰动的图片输入模型, 模型可以正确将其分类; 将加入特定扰动的图片输入模型, 模型将其错误分类。同时这种对抗样本的图像与正常图像在人眼看来并没有什么太大的变换。

与传统的机器学习模型存在的漏洞不同, 对抗样本的攻击并不是由于模型编写代码时的失误或是错误造成的。对抗样本是一种人为加入特定扰动的数据样本, 这些加入特殊扰动的数据样本在被输入到机器学习模型之后, 模型将其错误的分类, 但是这些特殊扰动是人类无法通过自身感知力察觉的。

目前, 已经有很多学者在对抗样本的图像领域做了研究。一方面, 一些研究人员将主要的扰动信息添加在了图片的重点区域, 以期望减少扰动信息, 做到更好的隐蔽性。但是图片的重点关注区域同样也是人眼的主要关注区域, 这些添加的扰动信息同样也容易被人眼所察觉, 隐蔽性也大打折扣。另一方面, 一些研究人员做到了扰动信息的隐蔽性, 添加这些扰动信息的图片很难被人眼所察觉, 但是生成这些对抗样本的图片所消耗的时间相对较长, 不利于大规模生成对抗样本。本文提出了一种聚焦图片边缘区域的攻击算法:

(1) 通过 YOLO 目标检测模型提取卷积神经网络对图像的重点关注区域;

(2) 以 MIFGSM 为基础, 结合 Curls 先梯度下降再梯度上升的思想, 对这些重点关注区域之外的区域进行扰动信息的添加;

(3) 为避免模型整体速度降低, 采用固定次数进行梯度下降优化, 最后形成局部扰动的图像对抗样本。

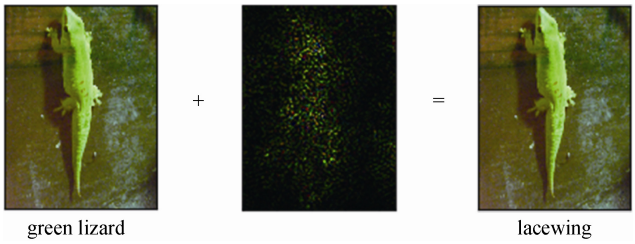

图 1 生成对抗攻击样本

Figure 1 Generates a counter attack sample

2 相关工作

在探索深度学习可解释性的研究过程中, Szegedy 等人[3]证明了深度学习模型对加入特定扰动的输入样本表现出极强的脆弱性, 并由此发现了对抗样本的存在,提出了第一个针对深度学习的对抗攻击方案 Box-constrained L-BFGS。Goodfellow 等人[4]设计了快速梯度对抗攻击的方法, 即 FGSM。他们认为由于深度神经网络存在高维线性特征, 导致高维线性模型存在对抗样本。通过调整修改幅度的大小来提高对抗样本攻击的成功率, 但是过大的修改幅度在提高攻击成功率的同时也容易被人眼所察觉。基于 FGSM 衍生出了 BIM 攻击方法[5]。Papernot 等人[6]提出了 JSMA 攻击, 不同于之前的添加全局扰动, JSMA 选择对抗显著值最大的像素进行扰动。Moosavi-Dezfooli 等人[7]提出了 DeepFool 方法, 通过每一次迭代, 沿着决策边界方向进行扰动, 逐步地将分类结果向决策边界另一侧移动, 使得分类器分类错误。随后 Moosavi-Dezfooli 等人[8]又提出了 Universal Perturbation 攻击, 不同于针对单张图片进行对抗攻击, Universal Perturbation 生成的扰动具有很强的泛化能力, 能够跨数据集、跨模型实施对抗攻击。One-Pixel 攻击通过仅改变原始图像中一个像素点实现针对深度神经网络的对抗攻击, 是一种基于前向传播的攻击方案[9]。Xiao 等人[10]提出以生成对抗网络为基础的对抗样本生成方法, 这种方法通过对抗网络(GAN)生成可用于攻击的对抗样本。Huang 等人[11]提出了一种中间层攻击的方法, 从而使对抗样本更具有迁移性。Han 等人[12]提出了一种多目标攻击的对抗样本生成网络, 可以将一个图片生成多种类型的对抗样本。Zhou 等人[13]提出了一种利用

GAN 来替代模型的方法, 这种方法不需要真实数据即可生成对抗样本。

对抗样本攻击不只单纯存在于计算机内部, 通过物理世界的设备和语音识别、自然语音处理方面也可以进行对抗样本攻击。Kurakin 等人[14]将对抗样本的图片打印出来后, 让手机摄像头对加入扰动的图片进行识别, 结果显示, 即使通过摄像头输入, 模型识别的结果仍然可能出现错误。Sharif 等人[15]将扰动加入到眼镜框上, 制作出一副特殊的对抗样本眼镜, 佩戴者戴上这副眼镜后, 人脸识别系统将其识别成其他人, 甚至出现性别识别错误的情况。Eykholt 等人[16]在交通标志上进行涂抹或贴图, 仅改变交通标志一小部分区域的像素, 导致汽车的识别系统产生判断错误, 从而对自动驾驶系统进行攻击。JunchengB 等人[17]在物理世界的摄像头上贴上一张特殊处理的透明图片, 当摄像头拍摄到物体时, 拍摄的图片会呈现出褪色或者眩光的效果, 这种方法依然可以产生对抗样本, 使模型识别错误。Brown 等人[18]将一张特制的图片放到真实物体旁边, 然后通过摄像头拍摄物体, 分类模型将其分为其他物体。Schonherr 等人[19]在语音中加入特殊噪声, 人耳识别时只会听到一些细微的电流声音, 但是语音识别系统对这种加入人为扰动的语音进行了错误的翻译。Jia 等人[20]在文本中加入几个多余字符, 使文本阅读理解系统产生错误回答。

图像分类是神经网络在计算机视觉方面最常见的应用, 神经网络的应用使得图像分类任务的效率得到了大幅提升。除了简单的进行分类任务, 在图像分类的基础上产生了更高级、更复杂的任务, 比如目标检测。目标检测被广泛应用在人脸识别、自动驾驶、工业检测、医学影像等领域。

由 Ross Girshick 等人[21]首次在目标检测任务上使用了卷积神经网络的方法, 提出了区域卷积神经网络(R-CNN), 这个模型取得了巨大的成功并影响深远。在此基础上, 又形成了 Fast R-CNN 模型, 相比于 R-CNN, 其计算量大大减少, 提高了处理速度, 并且引入了回归方法来调整目标物体的位置, 进一步提高了物体识别的准确性[22]。Joseph Redmon 等人[23]又提出 YOLO(You Only Look Once, YOLO)模型, 相比于 R-CNN, YOLO 使用单个网络结构, 且预测框少很多, 在速度上快的多, 能够达到实时响应的水平。经过不断地迭代升级, 至今 YOLO 模型仍有很强的竞争力[24]。

我们将局部扰动对抗样本与其它全局扰动和重点区域扰动添加对抗样本进行对比, 我们将对 FGSM、MIFGSM、PGD 全局扰动方法和 PS-MIFGSM 重点区域扰动方法进行介绍:

FGSM[4]: 该算法是 Goodfellow 等人提出的经典对抗样本算法, 是一种基于梯度生成对抗样本的快速梯度下降算法, 依据损失函数上升方向, 采用一个步长进行优化, 因此使对抗样本产生的速率极快。

MIFGSM[25]: 该算法是在 2018 年提出的对抗样本攻击算法, 与 FGSM 算法相比, MIFGSM 加入了动量因子, 具体表现形式如下:

$$x_{t+1}^{*} = x_{t}^{*} + a \cdot sign(g_{t+1})$$

$$g_{t+1} = u * g_{t} + \nabla_{x} J(x_{t}^{*}, y) / \| \nabla_{x} J(x_{t}^{*}, y) \|_{1}$$

其中损失函数 J, 真实样本 x, 真实样本类别 y, 扰动大小 μ, 迭代次数 T, 衰减系数 u。延梯度方向计算矢量更新 g_{t+1}, $sign$ 函数更新 x_{t+1}^{*}。

该算法是基于动量的迭代方法, 迭代的方向为梯度的反方向。同时在动量的迭代过程中为了稳定更新并且避免局部最值, 在损失函数的梯度方向上不断积累速度矢量, 以此提高生成对抗样本的攻击成功率。

PS-MIFGSM[26]: 该算法主要原理是借助 Grad-CAM 算法得到样本关注区域, 分为攻击区域和非攻击区域, 通过 MI-FGSM 计算得到干扰信息, 根据攻击与非攻击区域将扰动添加至原图像中, 该算法就是对图像的重点区域进行攻击。

PGD[27]: 该攻击是一次次的迭代攻击, 与 FGSM 的一步优化相比, 该算法更侧重于多次迭代, 能够在迭代中不断调整方向, 从而找出最优解。

受到上述方法对抗样本攻击和目标检测启发。本文提出了一种生成局部扰动对抗样本的方法。以 yolo_v3 作为目标检测模型, 先对图片进行目标检测, 对图片中物体的区域进行标记, 返回相应的位置信息。然后以 MIFGSM 为基础的对抗样本攻击方法, 对目标物体之外的区域进行扰动信息添加。同时, 为了提高对抗样本生成速度, 采用与 Curls 相似的方法, 先梯度下降再梯度上升, 越过决策边界, 生成对抗样本。

3 算法思路与算法设计

3.1 算法思路

相比于全局扰动, 局部扰动因为产生的扰动信息较少, 不容易被察觉。Liu 等人[28]证明在局部添加扰动的对抗样本攻击方法是有效的。然而大多数添加局部扰动的对抗样本扰动区域虽然相对较小, 但是与周围区域相比较依然容易观察。同样以图片为例, 添加局部扰动的区域与周围区域的对比度较高,

色彩差异较大, 因此在局部添加扰动生成对抗样本的工作仍需进一步完善。

在 Lu 等人[29]的论文中提到, 在测试对抗样本攻击效果的实验时, 如果将对抗样本图片中路标的背景裁剪掉, 路标被正确识别出的概率会大大提高。因此在被裁剪下的背景中, 存在着一定的扰动信息, 这些扰动信息对于图片的识别和分类起到一定的影响。受此启发, 如果在图片目标区域之外加入扰动信息, 同样也可以生成具有对抗攻击效果的对抗样本, 并且因为扰动信息在图片目标物体之外, 可以相对降低扰动信息被发现的概率。同时利用一些生成全局扰动的方法, 可以快速的生成扰动信息。因此, 最终设计思路是通过目标检测确定图片非重点区域, 结合其他对抗样本攻击生成扰动信息的方法, 形成局部扰动的对抗样本攻击方法。具体的方法如图 2 所示。通过目标检测部分确定添加扰动信息的位置, 然后再通过扰动生成部分生成扰动信息, 最后将扰动信息添加在合适的位置上, 生成最终的对抗样本。

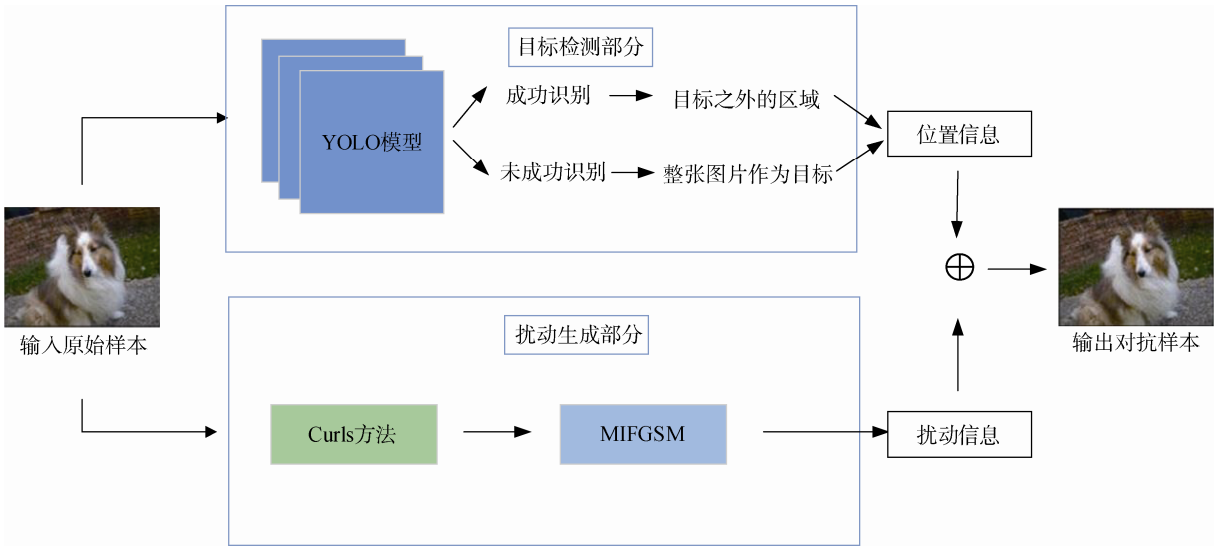

图 2　算法结构

Figure 2　Algorithm structure

3.2　位置信息检测

为了实现扰动的添加, 首先要获取添加扰动的具体位置信息。目标检测技术可以找出图片中的目标物体, 确定他们的位置和类别。得到图片中物体的位置信息后, 根据目标位置信息确定扰动信息添加的位置, 从而实现对抗样本的生成。所以使用目标检测是算法的第一步。

基于深度学习的目标检测模型分为两大类。一类是以 R-CNN 为代表的 Two Stage 的目标检测模型; 另一类是以 YOLO、SSD 为代表的 One Stage 的目标检测模型。Joseph Redmon 等人在 2015 年提出了 YOLO(You Only Look Once, YOLO)模型。YOLO 开创性地将候选区和目标识别两个阶段合二为一, 实现了端到端的目标检测。YOLO 先将整个图片划分为 $S*S$ 个方格区域, 如果一个物体落在一个方格内, 这个方格负责预测物体。方格对 B 个边界框进行置信度的预测, 置信度通过是否包含物体和交并比(IOU)计算。

$$Confidence = \Pr(Object) * IOU_{pred}^{truth}$$

其中边界框(bounding box)包括了边界框的中心位置 x 和 y, 边界框的宽度和高度 w 和 h。为了预测出物体的具体类别, 还需要再置信度的基础上乘以物体类别的概率, 最终的置信度预测为:

$$Confidence = \Pr(Class_i \mid Object) * \Pr(Object) * IOU_{pred}^{truth}$$

通过训练最终实现模型对目标物体的检测。使用 YOLO 模型对图片进行目标检测, 结果如图 3 所示。可以看出 YOLO 模型可以很好的将图片中的目标物体进行识别和分类。

在使用目标检测模型进行位置信息检测时, 并不需要模型返回的图片中物体的标签名称, 只需要物体的具体坐标, 即边界框左上、左下、右上、右下四个像素点的坐标。为了避免图片的缩放导致扰动信息错误的添加在其他位置, 将原有的四点具体坐标转化成相对位置信息, 从而可以适应图片缩放旋转等问题。由于目标检测模型不能达到 100%的识别准确率, 因此可能存在识别不到图片中物体的情况。可以通过调小识别置信度来提高识别准确率。同时, 对于识别不出目标位置的图片, 采用返回整张图片

位置信息的方法, 即将边界框左上、左下、右上、右下四个像素点设置为图片 4 个顶点。另外, 一些图片在进行目标检测时, 预测边界框的顶点坐标可能会出现在图片之外。对于这种超出图片本身的预测框, 将其顶点强制限制在图片内部, 避免出现错误。图 3 展示了目标检测部分如何得到重点区域位置信息。

图 3　YOLO 目标检测结果

Figure 3　YOLO target detection results

3.3　扰动信息添加

在确定扰动信息添加的位置后, 需要对样本进行扰动信息的添加。对抗样本攻击的方法有很多种, 包括局部扰动对抗样本和全局扰动对抗样本、特定目标攻击样本和非特定攻击目标样本、通用扰动攻击样本和非通用扰动攻击样本。常用的对抗样本攻击的方法包括 FGSM、BIM、MIFGSM、DeepFool 等方法。

MIFGSM 是 Dong 等人[30]提出的基于 FGSM 算法的改进, 该算法加入了动量因子。

MIFGSM 算法如下所示:

输入　分类模型 f 及损失函数 J, 真实样本 x, 真实样本类别 y, 扰动大小 μ, 迭代次数 T, 衰减系数 u。

输出　对抗样本 x^*

1)　$a=\mu/T$;

2)　$g_0=0$;　$x_0^*=x$;

3)　FOR $t=0$ TO $T-1$ DO;

4)　输入 x_t^* 到分类模型 f 中, 得到梯度 $\nabla_x J(x_t^*, y)$;

5)　延梯度方向计算矢量, 更新 g_{t+1}, 公式如下

$$g_{t+1}=u*g_t+\nabla_x J(x_t^*,y)/\|\nabla_x J(x_t^*,y)\|_1;$$

6)　用 sign 函数更新 x_{t+1}^*

$$x_{t+1}^*=x_t^*+a\cdot sign(g_{t+1});$$

7)　END FOR;

8)　RETURN　$x^*=x_T^*$

其中 1)中的 $a=\mu/T$ 是为了控制每次迭代的扰动大小, 避免每次迭代过程扰动过大。

通过 MIFGSM 算法可以看出, MIFGSM 的迭代方向为梯度相反的方向并且是基于动量的迭代方法。在此种动量迭代过程中在损失函数梯度方向上不断积累速度矢量来稳定更新并且避免局部最值, 提高生成对抗样本成功率。因 MIFGSM 方法的优良表现, 研究人员在此方法的基础上也提出了很多相近算法。

在 Shi 等人[31]的论文中提出了 Curls 方法。该方法认为, 如果直接使用迭代的方法, 图片向着梯度上升的方向形成一条迭代路径, 最终越过决策边界。这样的路径是单调的, 且不一定是最优的。可能还存在着更短的可以越过决策边界的迭代路径。因此, Curls 算法通过两条不同的路径寻找越过决策边界的最短路径。一条路径是传统的梯度向上的路径, 另一条是先梯度向下再梯度向上的路径。

Curls 算法如下:

输入　目标神经网络模型 $N(x)$, 可替代模型 $Sub(x)$, 真实样本 x, 真实样本类别 y, 初始噪音大小 ε, 迭代次数 T, 高斯扰动噪音方差 s, 步幅大小 α, 决策边界搜索步数 bs

输出　对抗样本 x^*

1)　初始化对抗样本平均方向 $\bar{R}$ 与两个起始点;

2)　$\bar{R}=0$,　$x_0^A=x$,　$x_0^B=x$;

3)　设置梯度下降标志 *downhill=Ture*;

4)　FOR $t=0$ TO T DO;

5)　$\xi_t^A, \xi_t^B \sim \boldsymbol{N}(0, s^2 I)$;

6)　计算梯度;

7)　$g_t^A=\nabla J_{sub}(x_t^A+\xi_t^A+\alpha\cdot\bar{R})$;

8)　$g_t^B=\nabla J_{sub}(x_t^B+\xi_t^B+\alpha\cdot\bar{R})$;

9)　IF *downhill=True* THEN

$$x_{t+1}^A=Clip_{x,\varepsilon}\{x_t^A-\alpha\cdot g_t^A\};$$

ELSE

$$x_{t+1}^A=Clip_{x,\varepsilon}\{x_t^A+\alpha\cdot g_t^A\};$$

END IF

10)　$x_{t+1}^B=Clip_{x,\varepsilon}\{x_t^B+\alpha\cdot g_t^B\}$;

11)　IF　*downhill=True*　*and*　$J(x_{t+1}^A)>J(x_t^A)$ THEN

downhill=False;

12) IF $N(x_{t+1}^A) \neq N(x)$ *or* $N(x_{t+1}^A) > J(x_t^A)$ THEN

UPDATE $\overline{R}$;

13) END FOR;

14) IF $N(x_T^A) \neq N(x)$ *or* $N(x_T^B) \neq N(x)$ THEN

15) IF $\| x_T^A - x \|_2 < \| x_T^B - x \|_2$ THEN

$x^* = x_T^A$;

ELSE

$x^* = x_T^B$;

对 x^* 进行二分搜索;

16) END IF;

17) RETURN x^*

其中第 12)中更新 $\overline{R}$ 是

$$\overline{R} = \frac{1}{K}\sum_{i=1}^{K} X^*, s.t.\ N(x^*)$$

这里的意思是记录并更新一张图片所有对抗样本的平均方向 $\overline{R}$, 并在第一步计算每一轮梯度时添加一个指向这个平均方向的向量。

而算法的第 14~16 步中提到的对 x^* 进行二分搜索实际上是指, 在 x 与 x^* 之间通过二分搜索的方式查找一个刚好越过决策边界的对抗样本。这样便可以在固定迭代长度中寻找更接近决策边界的对抗样本。

通过算法描述可以发现, Curls 算法是一种寻找最短扰动路径的方法。一条路径是沿着梯度向上的方向寻找对抗样本。另一条路径是先沿着梯度向下的方向寻找损失函数最低点, 然后再延梯度向上的方向寻找对抗样本。最后比较两条路径找到的对抗样本与原样本的距离大小, 选择最短距离(最小扰动)的对抗样本为最终的对抗样本。

受到Curls方法的启发, 在使用MIFGSM进行梯度上升添加扰动信息前, 先将图片沿着梯度下降的方向进行优化, 然后再沿着梯度上升的方向添加扰动信息。为了避免多次迭代的梯度下降优化导致模型整体速度降低, 可以采用固定的次数进行梯度下降优化。在保证模型整体速度的同时, 使模型生成的对抗样本的扰动信息更少。

对加入 Curls 方法的 MIFGSM 进行了测试。表 1 描述了加入 Curls 方法的 MIFGSM 和常规的 MIFGSM 之间的差别。

采用两种方法对 1000 张图片进行了测试, 在加入Curls方法的MIFGSM算法中, 前3次迭代均沿着梯度下降的方向进行的。从第 4 次迭代到最后结束, 迭代方向均沿着梯度上升的方向进行的。

表 1 加入 Curls 方法的对比

Table 1 Comparison of peer peer peer detection methods

攻击方法	迭代次数	时间	成功率(%)
MIFGSM	10	1min	93.83%
MIFGSM	13	1min7s	96.63%
Curls_MIFGSM	10	1min13s	94.04%
Curls_MIFGSM	13	1min26s	97.3%

通过对比容易观察到, 在迭代次数相同的情况下, 加入 Curls 方法的 MIFGSM 与常规的 MIFGSM 攻击成功率相比几乎一致。但是提高迭代次数后, 加入 Curls 方法的 MIFGSM 与常规 MIFGSM 迭代攻击次数相同时(不算延梯度下降的3次迭代), 加入Curls方法的攻击成功率更高。但是不排除因为总迭代次数增加导致的成功率提高。另外, 测试时发现加入 Curls 方法后, 生成对抗样本的时间增加了, 原因是强制进行了前 3 步的梯度下降操作。部分图片因为在一次或几次迭代后就可以越过决策边界, 变成对抗样本, 强制加入前 3 步后, 总体迭代次数增加, 最终导致总时间增加。

3.4 算法设计

前文分析了此方法在位置信息检测和对抗扰动信息生成两个部分的实现形式和方案选择。第 1 步通过目标检测模型获取图片中目标物体的位置信息, 第 2 步通过 MIFGSM 方法生成对抗扰动信息。最后将两步得到的结果相结合, 生成局部扰动的对抗样本。

算法描述如下:

输入 分类模型 f 及损失函数 J, 真实样本 x, 真实样本类别 y, 扰动大小 μ, 迭代次数 T, 衰减系数 u, 目标检测模型 L。

输出 对抗样本 x^{adv}

1) 输入 x 到 yolo_v3 中, 得到目标物体位置信息:

$$A^* = L[x]$$

2) 使用引入Curls方法的MIFGSM攻击 x, 得到原始对抗样本:

$$x^* = MIFGSM(x)$$

3) 将原始对抗样本目标物体位置设置为 0, 其余位置不变:

$$x^* = 0, ij \in A^*$$

4) 将原图 x 与步骤 3)修改过的对抗样本 x_{ij}^* 结合, 得到最终样本:

$$x^{adv} = x_{ij}^{*} + x$$

其中 2)中引入 Curls 方法的 MIFGSM 算法如下所示:

输入　分类模型 f 及损失函数 J, 真实样本 x, 真实样本类别 y, 扰动大小 μ, 迭代次数 T, 衰减系数 u。

输出　对抗样本 x^{*}

5) $a = \mu / T$;

6) $g_0 = 0$; $x_0^{*} = x$;

7) FOR $t=0$ TO T–1 DO;

输入 x_t^{*} 到分类模型 f 中, 得到梯度 $\nabla_x J(x_t^{*}, y)$;

延梯度方向计算矢量, 更新 g_{t+1}, 公式如下

$g_{t+1} = u \cdot g_i + \nabla_x J(x_t^{*}, y) / \| \nabla_x J(x_t^{*}, y) \|_1$;

8)　IF $t<3$ DO;

$x_{t+1}^{*} = x^{*} - a \cdot sign(g_{t+1})$

ELSE

$x_{t+1}^{*} = x^{*} + a \cdot sign(g_{t+1})$

9) END FOR;

10) RETURN　$x^{*} = x_T^{*}$

算法第 1 步为目标物体位置检测, 通过目标检测模型 L 返回图片中物体的位置信息。第 2 步为扰动信息的添加, 扰动信息的添加以 MIFGSM 为基础, 引入 Curls 先梯度下降再梯度上升的思想, 即算法的第 8 步。这里只在前三步进行梯度下降, 从第 3 步开始一直到最后延用传统的迭代梯度上升。第 3 步为清空目标物体位置之外位置信息, 被清空的位置将添加生成的扰动信息。最后一步是将对抗扰动信息添加在相应的位置上, 生成最终的对抗样本。

4　实验结果与分析

实验中本文所提出的算法以及实验对比均在百度的 AI Studio 平台上实现的, 内含的框架为百度的 PaddlePaddle 2.1.2 版本。

本文实验环境如表 2 所示。

表 2　实验环境描述

Table 2　Description of experimental environment

实验环境	实验配置
操作系统	Ubuntu 16.04.6
CPU	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz
GPU	NVIDIA Tesla V100
内存	16
编程语言	Python 3.7
深度学习框架	PaddlePaddle 2.1.2
开发环境	AI Studio

同时进行实验对比时, 算法参数保持一致, 其中部分参数设置如表 3 所示。

表 3　参数设置

Table 3　Parameter settings

参数	值
Num_size	1000
psilon	40
Num_iters	10
Decay_factor	1

其中:

Num_size 是指加入样本的多少, 在这里指的是测试选择的 1000 张作为对抗样本攻击的测试图片。

Epsilon 是指扰动大小, 为了确保对抗样本与真实样本之间差异不大, 扰动大小应选择合理。扰动大小越大, 对抗样本与真实样本差异越大。

Num_iters 是指迭代次数。FGSM 算法迭代次数为 1, 不存在多次迭代, 本文则迭代次数为 10。

Decay_factor 是指衰减因子。衰减因子只在 MIFGSM 和本文方法中出现。

使用的数据集为 ILSVRC2012 ImageNet 数据集。对比实验所使用的 vgg16, resnet50, googlenet 作为被攻击的对象同样是在 ImageNet 数据集上训练好的, 这三个被攻击模型作为 PaddlePaddle 自带的内置模型, 可以在测试时直接使用。实验过程中使用的原始图片是 ILSVRC2012 ImageNet 中选取的 1000 张图片。为了保证目标检测模型进行目标物体识别时可以返回较为准确的结果, 选择的 1000 张图片基本上是只包含一类物体, 并且图片不存在多个相同物体。同时, 为了避免被攻击样本在原始状态就分类错误, 对抗样本攻击的成功率是建立在原始图片识别正确的前提下。

本文提出的局部扰动的对抗样本攻击方法与 PGD、MIFGSM、FGSM、PS-MIFGSM 4 种对抗样本攻击方法进行了对比。用五种方法对 ImageNet 数据集中选出的 1000 张图片进行了对抗扰动信息的添加, 再用五种方法生成的对抗样本对 3 种图像分类模型进行了攻击。3 种图像分类模型分别是 vgg16, googlenet, resnet50。实验图片因为大小不同, 均进行了归一化的处理。像素值扰动的最大值为 40, 其中迭代攻击中, 迭代次数为 10, 衰减系数为 1。评估标准为对抗样本对图像分类模型攻击成功率。另外, 为了测试白盒攻击和黑盒攻击的差异, 以 vgg16 模型为白盒攻击测试模型, googlenet 和

resnet50 为黑盒攻击测试模型, 观察白盒攻击和黑盒攻击的攻击效果。

表 4 展示了几种对抗样本攻击方法的比较。表格的每一行表示的是 1 种攻击方法对于 3 种不同网络结构的图像分类模型的攻击效果。其中最后一行为本文提出的对抗样本攻击方法。3 种图像分类模型中, vgg16 为白盒攻击测试模型, googlenet 和 resnet50 为黑盒攻击测试模型。

表 4 攻击成功率对比

Table 4 Comparison of attack success rates (%)

攻击方法	VGG16	GoogleNet	ResNet50
FGSM	18.14	30.46	23.44
PGD	100	84.25	56.88
MIFGSM	100	31.35	24.11
PS-MIFGSM	100	34.25	30.58
本文	100	33.33	31.48

对于白盒攻击, 除了FGSM方法以外, 另外四种方法的攻击成功率都很高, PGD、MIFGSM 与 PS-MIFGSM 方法成功率达到 100%。本文提出的方法攻击的成功率也为 100%, 与 MIFGSM、PS-MIFGSM 成功率相同。在数据一致的情况下, 与同类型的PS-MIFGSM方法相比, 在攻击成功率与生成对抗样本时间上保持一致。而本文中提出的局部扰动方法避开了图像中人眼重点关注区域, 因此对抗扰动信息更具有隐蔽性, 不易被人眼所察觉。一方面证明了添加局部扰动的对抗样本一样可以达到和全局扰动对抗样本一样的攻击效果。另一方面也证明了图片中关键位置之外的区域也存在特征信息, 在图像分类模型训练时被模型学习, 从而使对抗扰动方法可以在这些区域添加扰动信息, 最终形成具有攻击性的对抗样本, 影响图片分类模型的分类效果。

对于黑盒攻击, 除了 PGD 方法攻击成功率较高以外, 另外 3 种方法的攻击成功率都不高, 在 30%左右。本文提出的方法在有些情况下甚至优于其余方法。

探究此种原因, 实验又在不同扰动大小的条件下进行了对比测试。将扰动大小 epsilon 的值分别调至 4 和 1 时发现, FGSM 的黑盒攻击成功率基本保持不变, 其他 4 种方法均有所下降, 但是本文方法与另外三种方法相比依旧较高。表 5 展示了在不同扰动大小下几种方法的黑盒攻击成功率。

通过表 5 所展示的不同情况下成功率对比, 可以看出本文方法在黑盒攻击中成功率明显下降, 主要原因有两个, 一是局部扰动的方法添加的扰动信息相对于全局扰动较少, 较少的扰动信息不易于产生攻击性, 因此无法达到很强的攻击性。二是因为不同的图像分类模型之间的网络结构有所差异, 分类的决策边界在不同模型间不尽相同。因此在白盒攻击越过决策边界的对抗样本不一定可以越过其他黑盒攻击模型的决策边界, 最终导致黑盒攻击效果较差。另外在几种方法生成的对抗样本图片中, MIFGSM 和 PGD 方法为了达到很高的攻击成功率, 在对抗样本上添加的扰动信息非常多, 以至于部分图片扰动信息可以肉眼识别到。特别是 PGD 方法加入的随机化操作, 使这些扰动信息更加明显。

除此之外, 还进行了对抗样本生成速度的测试。在未加入目标检测前本文提到的方法与 MIFGSM 方法生成单个对抗样本的时间基本一致, 加入目标检测后本文方法生成单个对抗样本的时间增加了 200 ms。综合比较下, 本文提出的方法与 PS-MIFGSM、MIFGSM 方法生成对抗样本时间基本一致, 可以快速的生成对抗样本。同时我们也比较了此方法与 DeepFool 方法生成对抗样本的速度。此方法速度是 DeepFool 方法速度的 4 倍左右。因此, 本文方法在生成对抗样本的速度上具有一定的优势。

表 5 不同扰动大小在黑盒攻击下的对比

Table 5 Comparison of different disturbance sizes under black box attack

扰动大小	攻击方法	GoogleNet(%)	ResNet50(%)
4	FGSM	31.25	18.75
4	PGD	21.62	10.81
4	MIFGSM	17.69	13.07
4	PS-FGSM	18.06	13.11
4	本文	17.94	13.67
1	FGSM	47.37	26.31
1	PGD	14.47	10.52
1	MIFGSM	17.39	11.31
1	PS-FGSM	17.65	11.21
1	本文	15.09	12.26

扰动区域与扰动大小也是一个重要的评价指标。本方法平均扰动区域的大小比 MIFGSM 方法扰动区域减少了 36%。同时对扰动大小进行了比较。如果对抗样本 x_* 攻击成功, 采用 L_2 距离来计算对抗样本 x_* 与原始样本 x 之间的扰动距离。比较时采用攻击成功的样本的平均值进行比较。表 6 展示了此方法与 PGD 和 MIFGSM 方法扰动区域与扰动大小的比较。

表 6　扰动区域与扰动距离对比

Table 6　Comparison of disturbance area and disturbance distance

攻击方法	扰动区域(%)	扰动距离
PGD	100	439
MIFGSM	100	35.14
本文	64	22.09

并且对四种攻击方法生成的攻击样本进行了比较, 如图四所示。其中第 1 行为未加入任何扰动信息的原始图片, 第 2 行为使用 FGSM 攻击方法生成的对抗样本, 第 3 行为使用 PGD 攻击方法生成的对抗样本, 第 4 行为使用 MIFGSM 攻击方法生成的对抗样本, 第 5 行为使用本文提出方法生成的对抗样本。其中通过 FGSM 和 PGD 方法生成的对抗样本相对于后两种方法生成的对抗样本, 图片中的噪点更明显, 更容易被人眼所发现。MIFGSM 与本文提出的方法在图片中产生的噪点, 从肉眼上来看, 更加趋于原图, 同时扰动有效。而本文所添加扰动信息区域相对于其他方法更少, 距离更短。同时实验发现, 因为本方法是在图片中目标物体之外的非重点区域加入的对抗扰动信息, 所以在对抗样本的分类结果上与其他几种方法生成的对抗样本分类结果有些许差异。这也说明了背景中加入的扰动信息与图片目标中加入的扰动信息对图片特征的影响有所差别。

5　总结与展望

本文提出了一种聚焦于图像重点区域之外的对抗样本攻击方法, 引入 YOLO 模型作为目标检测, 结合 MIFGSM 方法与 Curls 先梯度下降再梯度上升的思想, 在图像目标区域之外的其他区域生成扰动

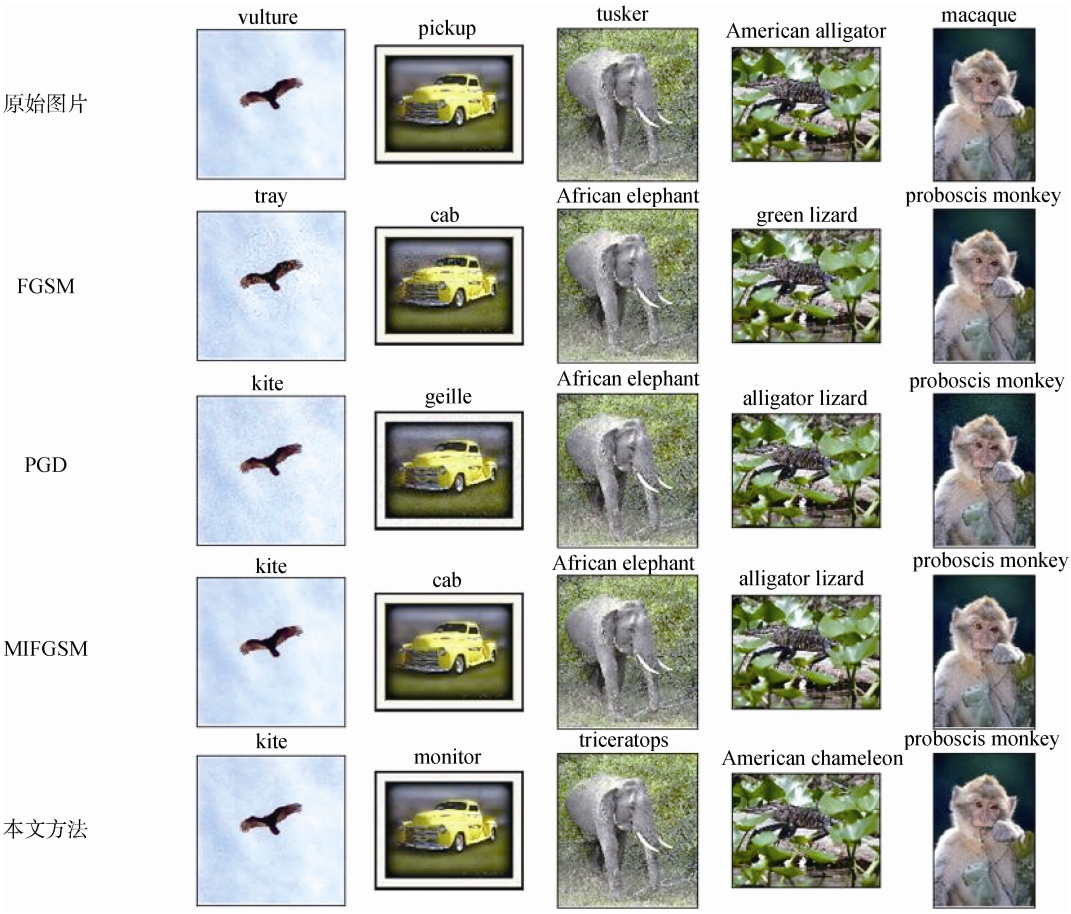

图 4　对抗样本的对比

Figure 4　Comparison of adversarial samples

信息, 从而实现对神经网络模型的攻击。在攻击模型的实验中发现, 该算法生成的对抗样本在白盒攻击下能保持很高的攻击成功率, 在黑盒攻击下与全局扰动的对抗攻击算法基本保持一致, 一些情况下甚至优于已有方法。

本文所提出的局部扰动对抗样本攻击方法在保证较高的攻击成功率的前提下, 通过减少图像的扰动区域实现对抗扰动信息的隐蔽性, 使对抗样本更接近真实样本。本文的方法在黑盒攻击中的表现相对较差, 下一步的工作我们将思考如何提高该方法的黑盒攻击成功率。

参考文献

[1] Liu H, Zhao B, Guo J B, et al. Survey on Adversarial Attacks towards Deep Learning[J]. *Journal of Cryptologic Research*, 2021, 8(2): 202-214.
(刘会, 赵波, 郭嘉宝, 等. 针对深度学习的对抗攻击综述[J]. *密码学报*, 2021, 8(2): 202-214.)

[2] Akhtar N, Mian A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey[J]. *IEEE Access*, 6: 14410-14430.

[3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. 2013: arXiv: 1312.6199. https://arxiv.org/abs/1312.6199

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572

[5] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]. *The 2017 ACM on Asia Conference on Computer and Communications Security*, 2017: 506-519.

[6] 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*, 2016: 372-387.

[7] Moosavi-Dezfooli S M, Fawzi A, Frossard P, et al. DeepFool: A simple and accurate method to fool deep neural networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[8] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[9] 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.

[10] Xiao C W, Li B, Zhu J Y, et al. Generating Adversarial Examples with Adversarial Networks[EB/OL]. 2018: arXiv: 1801.02610. https://arxiv.org/abs/1801.02610

[11] Huang Q, Katsman I, Gu Z Q, et al. Enhancing adversarial example transferability with an intermediate level attack[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 4732-4741.

[12] Han J F, Dong X Y, Zhang R M, et al. Once a MAN: Towards multi-target attack via learning multi-target adversarial network once[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 5157-5166.

[13] Zhou M Y, Wu J, Liu Y P, et al. DaST: data-free substitute training for adversarial attacks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 231-240.

[14] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]. Artificial Intelligence Safety and Security. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group. Chapman and Hall/CRC, 2018: 99-112.

[15] M Sharif,S Bhagavatula,L Bauer, et al. Accessorize to a crime[C]. *Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.

[16] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1625-1634.

[17] Li J C, Schmidt F R, Kolter J Z. Adversarial Camera Stickers: A Physical Camera-Based Attack on Deep Learning Systems[EB/OL]. 2019: arXiv: 1904.00759. https://arxiv.org/abs/1904.00759

[18] Brown T B, Mané D, Roy A, et al. Adversarial Patch[EB/OL]. 2017: arXiv: 1712.09665. https://arxiv.org/abs/1712.09665

[19] Schonherr L, Kohls K, Zeiler S, et al. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding[C]. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019: 210-226.

[20] Jia R, Liang P. Adversarial Examples for Evaluating Reading Comprehension Systems[EB/OL]. 2017: arXiv: 1707.07328. https://arxiv.org/abs/1707.07328

[21] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580-587.

[22] Bharati P, Pramanik A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey[M]. Computational Intelligence in Pattern Recognition. Singapore: Springer Singapore, 2019: 657-668.

[23] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779-788.

[24] Jiang P Y, Ergu D J, Liu F Y, et al. A Review of Yolo Algorithm Developments[J]. *Procedia Computer Science*, 2022, 199: 1066-1073.

[25] Liu Z H, Peng W Y, Zhou J, et al. MI-FGSM on Faster R-CNN Object Detector[C]. *ICVIP 2020*: *2020 The 4th International Conference on Video and Image Processing*, 2020: 27-32.

[26] Wu L R, Liu Z H, Zhang H, et al. PS-MIFGSM: Focus Image Adversarial Attack Algorithm[J]. *Journal of Computer Applications*, 2020, 40(5): 1348-1353.
(吴立人, 刘政浩, 张浩, 等. 聚焦图像对抗攻击算法PS-MIFGSM[J]. *计算机应用*, 2020, 40(5): 1348-1353.)

[27] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[OL]. 2019: ArXiv Preprint ArXiv:1706.06083

[28] Liu M X. *Local adversarial noise-based attack method and its defense for image classifier*[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.
(刘美汐. 基于局部噪声的图像分类器攻击与防御方法研究[D]. 北京: 北京邮电大学, 2020.)

[29] Lu J J, Sibai H, Fabry E, et al. NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles[EB/OL]. 2017: arXiv: 1707.03501. https://arxiv.org/abs/1707.03501

[30] Dong Y P, Liao F Z, Pang T Y, et al. Boosting adversarial attacks with momentum[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9185-9193.

[31] Shi Y C, Wang S Y, Han Y H, et al. Curls & whey: Boosting black-box adversarial attacks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 6512-6520.

王辛晨 于 2017 年在山东青年政治学院计算机科学与技术专业获得学士学位。现在大理大学计算机技术专业攻读硕士学位。研究领域为联邦学习。Email: 1553746204@qq.com

李晓伟 于 2013 年在西安电子科技大学信息安全专业获得博士学位。现任大理大学数学与计算机学院副教授。研究兴趣包括: 网络安全协议、云计算安全、区块链技术。Email: lixiaowei_xidian@163.com