软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(11):3512−3529 [doi: 10.13328/j.cnki.jos.006084]

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

一种基于进化策略和注意力机制的黑盒对抗攻击算法*

黄立峰[1,2], 庄文梓[1], 廖泳贤[1], 刘 宁[1,2]

[1](中山大学 计算机学院(软件学院),广东 广州 510006)

[2](广东省信息安全技术重点实验室,广东 广州 510006)

通讯作者: 刘宁, E-mail: liuning2@mail.sysu.edu.cn

摘 要: 深度神经网络在许多计算机视觉任务中都取得了优异的结果,并在不同领域中得到了广泛应用.然而研究发现,在面临对抗样本攻击时,深度神经网络表现得较为脆弱,严重威胁着各类系统的安全性.在现有的对抗样本攻击中,由于黑盒攻击具有模型不可知性质和查询限制等约束,更接近实际的攻击场景.但现有的黑盒攻击方法存在攻击效率较低与隐蔽性弱的缺陷,因此提出了一种基于进化策略的黑盒对抗攻击方法.该方法充分考虑了攻击过程中梯度更新方向的分布关系,自适应学习较优的搜索路径,提升攻击的效率.在成功攻击的基础上,结合注意力机制,基于类间激活热力图将扰动向量分组和压缩优化,减少在黑盒攻击过程中积累的冗余扰动,增强优化后的对抗样本的不可感知性.通过与其他 4 种最新的黑盒对抗攻击方法(AutoZOOM、QL-attack、FD-attak、D-based attack)在 7 种深度神经网络上进行对比,验证了该方法的有效性与鲁棒性.

关键词: 对抗样本;黑盒攻击;进化策略;注意力机制;压缩优化

中图法分类号: TP18

中文引用格式: 黄立峰,庄文梓,廖泳贤,刘宁.一种基于进化策略和注意力机制的黑盒对抗攻击算法.软件学报,2021,32(11):3512−3529. http://www.jos.org.cn/1000-9825/6084.htm

英文引用格式: Huang LF, Zhuang WZ, Liao YX, Liu N. Black-box adversarial attack method based on evolution strategy and attention mechanism. Ruan Jian Xue Bao/Journal of Software, 2021,32(11):3512−3529 (in Chinese). http://www.jos.org.cn/1000-9825/6084.htm

Black-box Adversarial Attack Method Based on Evolution Strategy and Attention Mechanism

HUANG Li-Feng[1,2], ZHUANG Wen-Zi[1], LIAO Yong-Xian[1], LIU Ning[1,2]

[1](School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China)

[2](Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China)

Abstract: Since deep neural networks (DNNs) have provided state-of-the-art results for different computer vision tasks, they are utilized as the basic backbones to be employed in many domains. Nevertheless, DNNs have been demonstrated to be vulnerable to adversarial attacks in recent researches, which will threaten the security of different DNN-based systems. Compared with white-box adversarial attacks, black-box attacks are more similar to the realistic scenarios under the constraints like lacking knowledge of model and limited queries. However, existing methods under black-box scenarios not only require a large amount of model queries, but also are perceptible from human vision system. To address these issues, this study proposes a novel method based on evolution strategy, which improves the attack performance by considering the inherent distribution of updated gradient direction. It helps the proposed method in sampling effective solutions with higher probabilities as well as learning better searching paths. In order to make generated adversarial example less perceptible and reduce the redundant perturbations after a successful attacking, the proposed method utilizes class activation mapping to group the perturbations by introducing the attention mechanism, and then compresses the noise group by group while ensure that the

* 基金项目: 国家自然科学基金(61772567); 中央高校基本科研业务费专项资金(19lgjc11)

Foundation item: National Natural Science Foundation of China (61772567); Fundamental Research Funds for the Central Universities (19lgjc11)

收稿时间: 2019-09-29; 修改时间: 2020-01-30, 2020-04-02; 采用时间: 2020-05-09

generated images can still fool the target model. Extensive experiments on seven DNNs with different structures suggest the superiority of the proposed method compared with the state-of-the-art black-box adversarial attack approaches (i.e., AutoZOOM, QL-attack, FD-attack, and D-based attack).

Key words: adversarial example; black-box attack; evolution strategy; attention mechanism; optimization of compression

随着深度学习技术的不断发展,深度神经网络(deep neural network,简称 DNN)在包括图像分类、物体识别、场景分割等多种计算机视觉任务中都获取了出色的表现[1−4].随着结构更复杂、层级数量更多的神经网络模型的出现(如 AlexNet[5]、VggNet[6]、InceptionNet[7]、ResNet[8]等),深度神经网络不仅在预测的准确度上获得了进一步的突破,也在不断拓广其实际的应用范围[9,10].

然而,深度神经网络在达到高性能的同时,也展现出面临对抗样本攻击的脆弱性,即恶意地对输入数据添加微小但难以察觉的扰动,将导致深度神经网络输出错误的结果.这种被恶意篡改的数据定义为对抗样本[11].在这种情况下,包括医学[12]、安防[13]、智能分析[14]等不同领域中,基于深度神经网络的应用系统都将面对这种潜在的威胁:Sharif 等人[15]通过将对抗样本图案打印至眼镜边框上来欺骗人脸识别系统;Athalye 等人[16]利用对抗攻击算法制造出在不同的光照和角度下欺骗分类器的 3D 打印物体;以涂鸦的方法对路牌上的图案进行简单的修改[17],就会导致无人驾驶系统无法正确识别路牌的类别;Lee 等人[18]利用对抗样本图案隐藏人体,这将对行人识别与跟踪系统产生威胁.因此,研究对抗样本的生成原理和算法实现,有助于分析基于深度学习的系统存在的安全漏洞,并建立相应的防范机制.

根据对抗样本的攻击场景设定,可以将对抗样本攻击方法分类为:(1) 白盒攻击,即攻击者可以获知被攻击目标模型的所有信息,包括训练集数据、神经网络结构、模型参数以及训练方式等[19−22];(2) 黑盒攻击,即神经网络相关的信息对攻击者来说是透明不可知的,攻击者只能通过提交输入数据并观察输出结果来进行交互,以此为基础生成对抗样本[23−29].

目前,大多数的攻击方法都是基于白盒场景下进行研究的.由于可以对目标模型的信息进行分析,因此这类方法大多是基于神经网络的反向传播与梯度下降算法来反向最大化模型的损失函数,生成可以误导神经网络的对抗样本.该类方法包括 FGSM[19]、BIM[20]、JSMA[21]与 C&W[22]等.

尽管白盒攻击理论上存在可行性,但在现实场景中,应用系统的网络结构和相关数据都是严格保密的,因此黑盒攻击比白盒攻击更接近实际的应用场景.攻击者只能观察到网络模型的预测结果,而且需要对交互查询的次数进行约束,面临更大的挑战.目前,黑盒攻击主要包含两类方法,其中,

- 一类黑盒攻击方法是基于迁移性的对抗攻击[23−25],通过在已知的替代网络模型上生成对抗样本,再迁移至目标模型,观察能否攻击成功.这种方法仅需要进行一次查询,但由于不同的神经网络模型结构千差万别,因此迁移攻击的成功率通常较低.
- 另一类是基于梯度拟合的黑盒攻击方法[26−30],即通过对目标模型多次交互查询来观察输出结果的变化,以此为基础近似估计网络模型的损失函数梯度方向.但是这类方法需要与模型往复查询多次,耗费大量的计算资源,效率较低;且拟合的梯度与真实的梯度存在差异,导致生成的对抗样本扰动幅度较大,难以应用在实际场景中.

对此,本文提出了一种基于进化策略和注意力机制的黑盒对抗攻击方法(如图 1 所示),主要由两部分构成.

(1) 基于协方差矩阵自适应进化策略的攻击方法.与传统方法从高斯分布或伯努利分布中采样向量的思路不同,本文方法充分考虑到攻击过程中损失函数梯度方向的分布关系,基于协方差矩阵迭代学习每次拟合的梯度方向信息,自适应更新较优的搜索路径,使采样的扰动向量主要在损失函数下降的窄谷方向上生成,以高几率采样到有效的扰动,减少与攻击模型交互查询的次数,提高黑盒攻击的计算效率与成功率(如图 1 中阶段 1 所示).

(2) 基于注意力机制的对抗样本压缩优化方法.由于黑盒攻击生成的对抗样本冗余信息较多,容易被人眼视觉系统所察觉,因此本文结合类间激活热力图方法对生成的扰动进行分组,并依次压缩优化,降低扰动幅度的大小.该方法主要考虑了神经网络的注意力机制与冗余扰动数据的内在联系,提升优化的

效果,在保持攻击效率的同时,增强了对抗样本的不可感知性(如图 1 中阶段 2 所示).

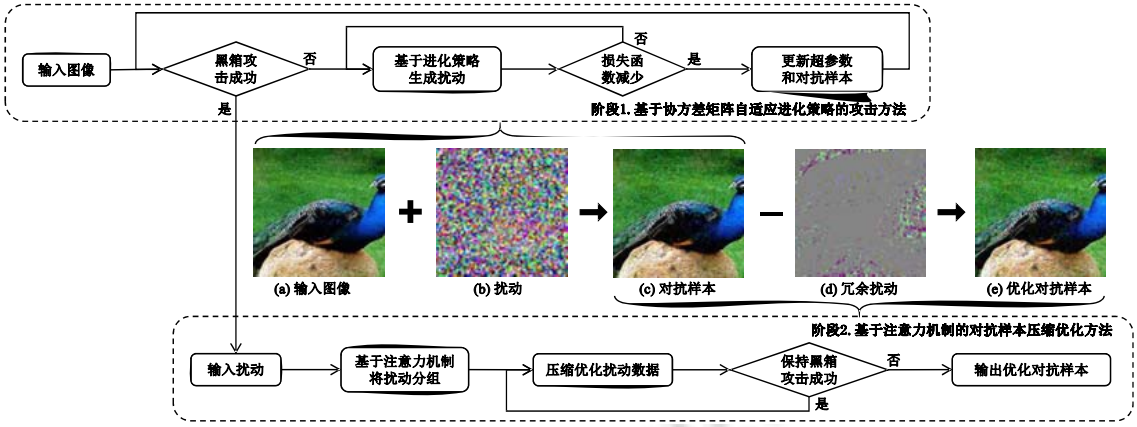


Fig.1 Pipeline of the proposed black-box adversarial attack method

图 1 本文提出的黑盒对抗攻击方法流程图

本文的主要贡献分为 3 个方面.

(1) 提出了基于协方差矩阵自适应进化策略的攻击方法.该方法考虑了梯度方向分布的特性,可以较大地提升黑盒攻击的效率,有较强的实用性.

(2) 提出了基于注意力机制的对抗样本压缩优化方法,结合类间激活热力图对扰动进行优化,可以有效地减少冗余的数据,增强对抗样本的隐蔽性.

(3) 分析了注意力机制与对抗样本的内在关联,验证了本文方法的可靠性.

本文第 1 节介绍相关工作.第 2 节详细描述本文提出的基于进化策略和注意力机制的黑盒对抗攻击方法.第 3 节为实验设置与结果分析.最后一节为本文总结.

1 相关工作

1.1 对抗样本与白盒攻击方法

Szegedy 等人[11]首次提出了对抗样本的概念:对于分类任务,给定已训练好的深度神经网络 f,输入合理的图像数据 x,可以计算得到正确的结果 y,定义为映射函数 $y=f(x)$.而对抗攻击的目的是找到一个微小的扰动向量 δ,求解以下优化问题:

$$y'=f(x+\delta)\ \text{s.t.}\ \|\delta\|_p\leqslant\varepsilon, y'\neq y \tag{1}$$

其中,$\hat{x}=x+\delta$ 定义为对抗样本,误导深度神经网络输出错误的结果 y'.为了满足扰动的不可感知性,扰动向量 δ 需要满足 L_p 范数约束.对于非目标攻击,攻击者需要最小化正确结果 y 的概率,直到输出结果 $y'\neq y$;对于目标攻击,则需要最大化目标结果 $y'=\tilde{y}$ 的输出概率.

白盒场景中,攻击者可以获知目标模型的训练集数据、神经网络结构、模型参数等信息,通常利用神经网络模型的反向传播特性来生成对抗样本,可以分类为:(1) 单步攻击,包括 Goodfellow 等人[19]提出的 FGSM 方法和 Szegedy 等人[11]提出的 L-BFGS 方法;(2) 迭代攻击,包括 Moosavi-Dezfooli 等人提出的 DeepFool 方法[31]和 UAP 方法[32]、Kurakin 等人[20]提出的 BIM 算法以及 Carlini 等人[22]提出的 C&W 方法等.

1.2 黑盒攻击方法

由于深度神经网络内部信息的不可知性约束,因此黑盒对抗攻击面临更大的挑战,通常可以分为 3 类.

(1) 基于网络迁移性的攻击:该类方法假设不同的深度神经网络在高维空间中具有相似的分类边界,因此首先通过在已知的白盒模型中生成对抗样本,进一步迁移攻击未知的黑盒模型.Papernot 等人[23]利用

目标模型的数据集训练一个新的替代模型,并利用替代模型为白盒场景生成对抗样本欺骗目标模型;另一些研究[24,25]则专注于研究对抗样本在模型之间的迁移能力.与传统方法主要攻击神经网络输出层的思路不同,Zhou 等人[25]选择对中间特征层进行攻击,通过使范数距离最大化,可以较好地提升黑盒场景下的迁移成功率.Dong 等人[24]基于神经网络的平移不变性质,将替代模型计算得到的梯度向量进行卷积操作,成功地让对抗样本躲避了大多数防御方法的检测.

(2) 基于决策反馈的攻击:该类方法约束模型只反馈预测的标签结果,目前相关的研究较少,主要思路是将对抗样本作为初始数据来逐步逼近原始图像,在保持对抗攻击成功的前提下,不断缩小二者的距离,直到满足攻击成功的条件.这类方法需要预设定对抗样本进行初始化,对超参数不敏感,有较强的实用性.Brendel 等人[33]提出的 D-based Attack 方法通过在决策边缘寻找距离更小的对抗样本来进行攻击.Dong 等人[34]提出了基于进化策略的 Face Attack 方法,该方法不依赖替代模型,也不需要对梯度方向进行拟合,而是每次从分布中采集一个样本对模型的分类边界进行几何建模,并以此为基础判定对抗样本的移动方向,并自适应调整步长和分布参数,对当前数据进行更新,直到满足最大迭代次数则停止.

(3) 基于概率反馈的攻击:这类方法约束模型能输出标签的概率信息,通过与目标模型进行交互查询,观察输出结果的置信度数值变化来拟合损失函数的梯度方向,并结合迭代攻击的思路生成对抗样本.

本小节对几种主流的梯度拟合攻击方法进行简单介绍.

Bhagoji 等人[26]提出了基于有限差分的黑盒攻击方法(finite differences based method,简称 FD-Attack),该方法首先将图像的像素进行分组,然后依次修改各组的像素数据来近似估计每一组的梯度方向,最后根据所有分组的拟合梯度进行迭代攻击,在 MNIST 和 CIFAR-10 数据集上取得了较好的效果.其中,每一组像素的梯度计算公式如下所示:

$$g_i \approx \left[\frac{f(x+\delta e_i)-f(x-\delta e_i)}{2\delta}\right] \tag{2}$$

Chen 等人[27]提出了基于零阶优化的攻击方法(zeroth order optimization based attacks,简称 ZOO),该方法每次迭代只修改一个像素的数据来观测损失函数的变化,在获取近似梯度的基础上,结合 Hessian 矩阵与二阶牛顿法优化求解更精确的梯度来进行攻击.同时,该方法提出了分层式攻击策略(hierarchical attack),用于泛化攻击高分辨率的图像.其中,Hessian 矩阵的计算公式如下所示:

$$H \approx \frac{f(x+\delta e)-2f(x)+f(x-\delta e)}{\delta^2} \tag{3}$$

在 ZOO 方法[27]的基础上,Tu 等人[28]提出了基于自动编码器的零阶优化方法(autoencoder-based zeroth order optimization method,简称 AutoZOOM).该方法采用自动编码器降低搜索空间的维度,同时生成单位长度为 1 的随机向量来估计损失函数的梯度方向,提升黑盒攻击的效率.在 MNIST、CIFAR-10 和 ImageNet 的数据集实验中,AutoZOOM 展现出比 ZOO 方法更优的效果.其中,该方法的梯度拟合公式如下所示:

$$g \approx \frac{1}{n}\sum_{i}^{n} b\cdot\frac{f(x+\beta u_i)-f(x)}{\beta}\cdot u_i \tag{4}$$

Ilyas 等人[29]基于自然进化策略算法提出了查询限制攻击方法(query-limited method,简称 QL-Attack),该方法假设攻击者与目标模型的交互次数是有限的,并且查询的次数与攻击者付出的代价(如经济花费、计算开销等)成正比.主要思路是:通过在标准的正态分布中随机采样向量作为局部扰动,并统计损失函数的数值变化来计算模型的梯度方向,如公式(5)所示.

$$g \approx \nabla_x \boldsymbol{E}_{\mathcal{N}(z|x,\sigma^2)} f(z) \tag{5}$$

此外,Su 等人[30]基于差分进化算法提出了单像素攻击方法(one pixel attack),探索了一种极限条件下的攻击模式,即,仅修改图像中的一个像素来欺骗分类器.每次迭代中,攻击者基于父样本生成大量只修改一个像素的子样本,然后从中选择效果最优的结果来更新对抗样本.该方法不需要拟合损失函数的梯度方向,生成的扰动噪

点很少,但相对明显.此外,由于限制了修改像素的数量,因此在攻击高分辨率图像时,无法获得较好的效果.

然而,上述方法在迭代过程中都需要进行大量的交互查询.由于没有探索梯度分布的内在联系,每次更新都需要通过随机采样来拟合损失函数的梯度方向,因此效率相对较低,难以实际应用.针对于此,本文基于自适应进化策略,提出了一种高效的黑盒攻击方法,利用协方差矩阵迭代学习损失函数梯度方向的分布信息,寻找较优的搜索路径,更有效地生成扰动向量进行攻击.

1.3 注意力机制

Zhou 等人[35]首先发现:深度神经网络不仅在图像分类任务上有着优秀的表现,同时还能判定图像中的关键区域,并将量化的结果定义为类间激活热力图(class activation mapping,简称 CAM).该方法先对神经网络卷积计算得到的特征层进行加权求和,再上采样到图像空间,即可观察到神经网络的关注区域.Ramprasaath 等人[36]提出了基于梯度的类间激活热力图方法(gradient-based class activation mapping,简称 grad-CAM),将权重的计算方式更改为梯度的全局平均值,更灵活地应用于各类网络结构.

考虑到图像关键区域对神经网络的决策有着重要的影响,Liu 等人[37]提出了基于视觉敏感机制的对抗攻击方法(perceptual-sensitive GAN,简称 PS-GAN).该方法结合注意力机制,计算目标模型对图像最关注的一小块区域,然后利用生成对抗网络在该区域内生成扰动图案进行攻击,同时加入优化函数进行约束,减小生成图案与背景的差异,降低视觉的敏感性,成功欺骗了多种类型的神经网络.与 PS-GAN 方法限制扰动区域的思路不同,本文方法重点探索注意力机制与扰动密度分布的内在联系,先按神经网络的关注程度将图像进行分组,再依次压缩各组的冗余信息,从整体上减小对抗样本的扰动幅度,增强对抗样本隐蔽性.

2 基于进化策略的黑盒对抗攻击方法

本节对基于进化策略和注意力机制的黑盒对抗攻击算法进行介绍.其中,第 2.1 节详细阐述基于协方差矩阵自适应进化策略的黑盒攻击方法的原理;第 2.2 节针对相关黑盒攻击中扰动幅度较大的缺陷,介绍基于注意力机制的对抗样本压缩优化方法.

2.1 基于协方差矩阵自适应进化策略的攻击方法

由于深度神经网络的设计结构和模型参数的不可知性约束,攻击者无法直接利用梯度信息来生成对抗样本.本文提出了基于协方差矩阵自适应进化策略[38]的攻击方法(covariance matrix adaptation evolution strategy based adversarial attack method,简称 ES-Attack),可以在黑盒场景下高效率地拟合模型梯度方向进行攻击.

黑盒对抗攻击方法 ES-Attack 的流程如算法 1 所示.

算法 1. 基于协方差矩阵自适应进化策略的攻击方法.

输入:深度神经网络模型 f,损失函数 L,图像 $x\in\boldsymbol{R}^n$,最大迭代次数 T,采样数量 M,目标标签 $\tilde{y}$,标准差参数 σ,扰动幅度 ε,步长 α.

输出:扰动数据 δ,对抗样本 $\hat{x}$.

初始化 $x_0\leftarrow x$,搜索路径 $p_0\leftarrow\boldsymbol{0}$,协方差矩阵 $C_0\leftarrow\boldsymbol{I}^m$,向量集合 $U\leftarrow\varnothing$,$Z\leftarrow\varnothing$.

1. **while** $t\leqslant T$ and $f(\hat{x})\neq\tilde{y}$ **do**
2. $U\leftarrow\varnothing$, $Z\leftarrow\varnothing$
3. **for** i=1 to M **do**
4. $z_i\sim\mathcal{N}(0,\sigma^2C_t)$ //基于协方差矩阵随机采样
5. $u_i\leftarrow Upsample(z_i,\boldsymbol{R}^n)$ //将向量 z_i 上采样至 $\boldsymbol{R}^n$ 空间
6. $U\leftarrow U\cup u_i$, $Z\leftarrow Z\cup z_i$
7. **end for**
8. $g_t\leftarrow Gradient(x_t,U,\tilde{y})$ //估计损失函数的梯度方向
9. $\hat{x}\leftarrow Clip_{x,\varepsilon}(x_t+\alpha\cdot g_t/\|g_t\|)$ //结合约束更新对抗样本

10. **if** $L(\hat{x}) < L(x_t)$ **then**
11. $x_{t+1} \leftarrow \hat{x}, \delta \leftarrow \hat{x} - x$
12. $p_{t+1}, C_{t+1} \leftarrow Update(Z, p_t, C_t)$ //更新搜索路径与协方差矩阵
13. **end if**
14. **end while**
15. **return** $\delta, \hat{x}$

每次迭代中,主要包含以下两个步骤:

(1) 梯度拟合阶段(算法 1 中第 2 行~第 9 行).

该阶段基于协方差矩阵对损失函数的梯度方向进行拟合计算:首先,根据当前协方差矩阵从分布$\mathcal{N}(0,\sigma^2 C_t)$中采样若干扰动向量,以此为基础生成偏移向量;然后修改当前对抗样本,产生一组候选样本,通过统计候选样本的损失函数变化值估计梯度方向;最后将拟合的梯度进行归一化处理,结合扰动幅值约束更新对抗样本 $\hat{x}$.

为了提高计算效率,受到 ZOO 方法中分层式攻击策略(hierarchical attack)[27,34]的启发,ES-Attack 先从低维空间 $\boldsymbol{R}^m$ 中随机采样 M 个扰动向量 z 至集合 Z,再基于双线性插值方法将低维向量 z 上采样到图像空间 $\boldsymbol{R}^n$ 中,最后得到偏移向量 u.关于维度 m 的选择,本文将在第 3.2 节进行讨论.

拟合的梯度方向 g_t 描述了损失函数 L 在不同方向上的变化趋势,由偏移向量集合 U 中所有的元素加权平均计算得到.在 ES-Attack 中,各偏移向量的权重为损失函数的变化量,即让 L 减少更多的向量在评估梯度时占有更重要的地位,计算公式如下:

$$\Delta L(x_t, u_i, \tilde{y}) = L(x_t + u_i, \tilde{y}) - L(x_t - u_i, \tilde{y}) \tag{6}$$

$$g_t = -\frac{1}{M}\sum_{i=1}^{M} u_i \cdot \Delta L(x_t, u_i, \tilde{y}), u_i \in U \tag{7}$$

其中,本文方法选择使用交叉熵(cross entropy)作为损失函数 L.对于非目标攻击,$L(x,y)=\log f(x)_y$,当 L 减小时,f 输出正确结果 y 的概率也随之减小.对于目标攻击,$L(x,\tilde{y}) = -\log f(x)_{\tilde{y}}$,$L$ 减小意味着模型 f 输出指定目标标签 $\tilde{y}$ 的概率增加,本文主要考虑目标攻击.如算法 1 中第 9 行所示,本文选择使用 L_2 攻击生成对抗样本,若替换为 $\hat{x} \leftarrow Clip_{x,\varepsilon}(x_t + \alpha \cdot \mathrm{sign}(g_t))$ 则转变为 L_∞攻击,其中,sign 为符号函数.

(2) 参数更新阶段(算法 1 中第 10 行~第 13 行).

若更新的对抗样本没有使损失函数减小,则返回步骤(1)梯度拟合阶段,重新采样扰动向量并估计梯度方向;否则视为一次成功的搜索,并基于该次采样的向量数据对扰动和相关参数依次进行更新.在第 t 次迭代过程中,搜索路径与协方差矩阵参数的更新规则定义如下:

$$p_{t+1} = (1-c)p_t + \frac{\sqrt{c(2-c)}}{\sigma}\left(\frac{1}{M}\sum_{i=1}^{M} z_i\right), z_i \in Z \tag{8}$$

$$C_{t+1} = (1-c_1-c_2)C_t + c_1(p_{t+1}^T \cdot p_{t+1}) + c_2\left(\frac{1}{M}\sum_{i=1}^{M} z_i^T z_i\right), z_i \in Z \tag{9}$$

其中,c,c_1 和 c_2 分别为搜索路径和协方差矩阵的学习率.更新的路径 p_{t+1} 描述了下一次随机采样的分布均值的移动方向,在该次迭代搜索路径 p_t 的基础上,通过对向量集合 Z 中的所有元素平均求和计算得到.这类似于 MI-FGSM 方法[39]中动量迭代的思想,使向量 z 中相反方向的分量互相抵消,同时叠加相同方向的分量,确保搜索路径沿着损失函数的梯度方向移动.协方差矩阵 C_{t+1} 的更新公式包含了 3 部分,分别是已经学习到的知识 C_t、下一次迭代的搜索路径信息 p_{t+1} 以及满足当前损失函数下降的采样向量集合 Z.一种直观的解释是:协方差矩阵会不断学习历史过程中成功的搜索路径信息,并使下一次随机采样的向量以更大的几率分布在让损失函数减小的方向上.当 ES-Attack 的迭代次数达到预设的上限 T 或攻击成功时,停止黑盒攻击,并返回当前生成的扰动数据 δ 和对抗样本 $\hat{x}$.

相比之前基于梯度拟合的黑盒攻击方法,本文提出的 ES-Attack 方法有两个明显优势.

(1) 有效减少梯度拟合过程的交互查询次数.在 FD-Attack 方法[26]与 ZOO 方法[27]中,拟合梯度方向所需的查询次数依赖于图像维度的大小.若输入的图像维度为 n(如对于 Inception 模型 n=299×299×3),每次通

过修改 q 维数据来估计目标模型 f 的偏导梯度方向,则每一轮迭代需要 $O(n/q)$次查询计算.而在 ES-Attack 方法中,拟合梯度的查询次数与图像维度无关,每一轮迭代需要进行 $O(M)$次查询来估计梯度方向,其中,M 为算法中随机采样的个数.对于 ImageNet 数据集这类高维度图像来说,通常情况下满足 $O(n/q)>>O(M)$.因此 ES-Attack 以较少的交互查询次数来估计梯度方向,有效提升黑盒攻击的效率.

(2) 具有更大的概率采样到较优的扰动向量.对于图像数据,它的对抗样本更可能在特征空间中大致相同的梯度方向上进行分布[40].因此,相比于基于标准正态分布条件下随机采样计算梯度的 QL-Attack 方法[29],本文提出的 ES-Attack 考虑了迭代过程中梯度更新方向连续分布的特性,通过自适应方式学习成功搜索的历史路径信息,并以此为基础更新协方差矩阵,调整下一次搜索方向,使采样生成的偏移向量有更大的概率向损失函数减小的方向进行移动.直观上看,该方法通过加强每次迭代中梯度更新方向的内在联系来增加采样到较优扰动的概率.

2.2 基于注意力机制的对抗样本压缩优化方法

在黑盒场景中,损失函数的梯度方向是通过多次观察神经网络模型反馈的置信度数值变化进行拟合得到的,因此在每一次迭代过程中,拟合的梯度方向与真实的梯度方向必然存在着一定的误差,而偏离真实梯度方向的部分则称为冗余梯度.随着迭代次数的增加,冗余的梯度会不断积累,导致最终生成的扰动幅度较大[40].因此,本文提出一种基于注意力机制的对抗样本压缩优化方法(attention mechanism based compression method,简称 AM-Com),减少生成的冗余信息,降低对抗样本的扰动幅度.Zhou 等人[35]提出了类间激活热力图的概念,用于量化深度神经网络对图像内不同区域的关注程度.根据观察发现,热力图展现出以下两个特性.

(1) 不同的神经网络对于一张图像计算得出相似的热力图.尽管神经网络之间的设计结构和模型参数不同,但都基于图片中类似的区域进行决策计算.如图 2 所示,所有的模型之间共享着相似的高关注区域(动物的头部)和低关注区域(图像背景),其中,对抗防御模型的高关注区域分布相对较窄,且处于预训练模型的高关注区域内.在观察的基础上,本文对不同模型关注区域的相似程度进行了量化,通过计算两张热力图数值之差小于 0.1 的区域所占图像整体的比例,来衡量模型之间热力图的相似度.本文随机选取了 1 000 张图像对热力图的平均相似度进行统计,并将结果展示于图 3.通过对比可以看出,即使是关注区域差异最大的两个模型(即 VggNet-16 与 Inception-v3$_{ens4}$),平均也有超过 60%的区域的关注程度是非常相近的.这种模型间共享关注区域的特性验证了 Dong 等人[24]的结论,也为 AM-Com 方法提供了基础.

(2) 热力图的数值与对抗样本的扰动幅度密切相关.经实验发现:对分类结果影响重要的区域,生成扰动的密集程度也相对更大;另一方面,扰动的冗余信息也表现出相似的区域性特征,即高关注度区域和低关注度区域存在更多的扰动冗余(本文将在第 3.3 节进行讨论).因此,可以通过对不同区域的数据分别进行压缩优化,减少最终生成扰动的幅度,增加对抗样本的隐蔽性.

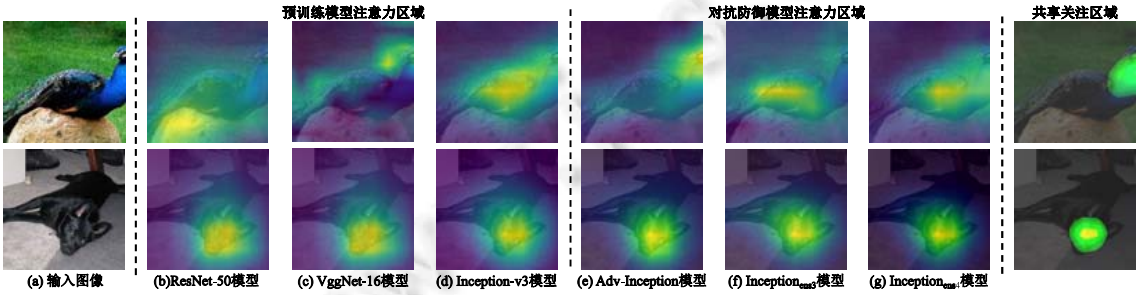


Fig.2 Attention regions between neural network models (The attention regions of models are highlighted)

图 2 神经网络模型的注意力区域(高亮部分为模型的关注区域)

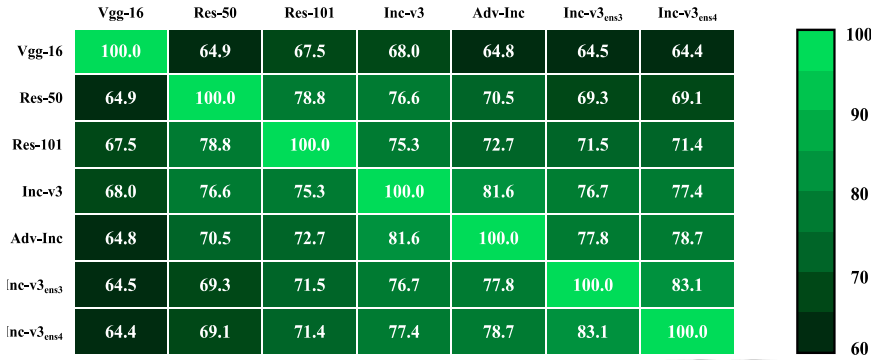


Fig.3 Similarity of heatmaps between different models(%)

图 3 不同模型生成热力图的相似度(%)

基于以上两个特性,本文提出的 AM-Com 方法主要包含 3 个步骤:(1) 基于替代模型生成热力图,近似估计目标模型的关注区域;(2) 根据冗余扰动数据的区域性特征,按模型的关注程度对扰动进行分组;(3) 依次压缩各组的扰动数值,直到所有分组的扰动数据完成优化步骤.流程如算法 2 所示.

算法 2. 基于注意力机制的对抗样本压缩优化方法.

输入:目标网络模型 f,替代网络模型 f_s,初始扰动向量 δ,图像 $x\in\boldsymbol{R}^n$,目标标签 $\tilde{y}$,优化系数 r_0,$r_{\max}$.

输出:优化后的对抗样本 $\tilde{x}$,扰动 δ'.

1. $h_s\leftarrow Normalize(CAM(f_s,x))$ //生成热力图并归一化
2. $h\leftarrow ReSample(h_s,\boldsymbol{R}^n)$ //将热力图重采样至 $\boldsymbol{R}^n$ 空间
3. $\delta'\leftarrow Group(\delta,h,N)$ //根据热力图对扰动分组
4. **for** i=1 to N **do**
5. $s\leftarrow r_0$
6. **while** $f(x+\delta')=\tilde{y}$ **do**
7. $\delta'[i]\leftarrow s\cdot\delta'[i]$ //压缩第 i 组扰动
8. $s\leftarrow\min(1.1\times s,r_{\max})$ //调整优化系数
9. **end while**
10. $\delta'[i]\leftarrow\delta'[i]\div s$
11. **end for**
12. $\tilde{x}\leftarrow x+\delta'$
13. **return** δ', $\tilde{x}$

算法 2 的具体描述如下.

(1) 热力图生成阶段(算法 2 中第 1 行、第 2 行).攻击者先基于替代模型 f_s 计算输入图像 x 的归一化热力图 h_s,然后通过双线性插值方法将热力图重采样到目标模型的输入空间中,获得数据维度为 R^n 的热力图 h,用于近似估计目标模型 f 对图像各区域的关注程度.

(2) 扰动分组阶段(算法 2 中第 3 行).该阶段根据热力图 h 将扰动 δ 进行像素级分组.首先将热力图的值域平均划分为 N 个区间,再根据下标索引将扰动的像素与热力值一一对应,最后按照热力值的所属区间范围将扰动的每一个像素分组至对应的集合.该扰动分组方法定义如下:

$$\delta'[i]=\left\{\delta_k\left|\forall k,h_k\in\left(\frac{i-1}{N},\frac{i}{N}\right]\right.\right\},\ i=1,2,3,...,N;\ k=1,2,3,...,n \tag{10}$$

其中,$\delta'[i]$是热力值属于第 i 个区间的扰动像素的集合,k 是分配于第 i 组的像素下标.i 的值越大,意味着目标模型对该组区域的关注程度越高.关于分组数量参数 N 的选择,本文将在第 3.3 节进行讨论.

(3) 冗余压缩阶段(算法 2 中第 4 行~第 11 行).AM-Com 方法按升序对分组后的扰动分别进行优化.对于

每一集合的扰动数据$\delta'[i]$,初始优化系数为r_0.在保持攻击成功的前提下,对集合内的数据进行压缩,并在迭代的过程中逐步增大优化系数.直到扰动δ'无法攻击成功,则停止压缩该集合的扰动数据.当所有集合都经过该优化过程,返回更新后的扰动和对抗样本.

3 实验评估

本节通过在 ImageNet 数据集[41]上实验来验证本文提出的方法在黑盒攻击场景下的有效性.主要包含以下内容:第 3.1 节介绍实验设置的相关内容;第 3.2 节讨论 ES-Attack 攻击方法相关的参数设置实验;第 3.3 节为 AM-Com 扰动压缩优化方法的相关实验;第 3.4 节将本文提出的方法与 4 种相关工作进行对比;第 3.5 节为对抗训练实验.

3.1 实验设置

本文从 ImageNet 数据集中随机选择 1 000 张图像作为实验测试数据.相比于非目标攻击,目标攻击更具有难度与挑战性,因此所有的实验都是基于目标攻击进行的.为了公平地对比,每张测试图像的攻击标签$\tilde{y}$将随机选择,并在实验过程中保持不变.

为了探讨不同模型之间的内在特性,本文将基于 7 种不同结构的神经网络模型进行讨论,包括 4 种预训练模型 VggNet-16[6]、InceptionNet-v3[7]、ResNet-50[8]和 ResNet-101[8]以及 3 种通过对抗训练[42]得到的鲁棒防御模型 Adv-Inception、Inception-v3$_{ens3}$ 和 Inception-v3$_{ens4}$.以下实验将这 7 种网络模型简称为 Vgg-16、Inc-v3、Res-50、Res-101、Adv-Inc、Inc-v3$_{ens3}$ 和 Inc-v3$_{ens4}$.在第 3.2 节与第 3.3 节的实验中,本文主要基于广泛部署于实际场景的 4 种预训练模型来寻找提出方法的最优参数,且在第 3.4 节的实验中加入 3 种对抗防御模型进行测试,验证提出方法的泛化性和鲁棒性.

实验的评估指标包括:(1) 交互查询次数(query numbers,简称 QN)[28,29],即攻击过程中查询目标模型输出结果的次数,次数越少,则代表攻击效率越高;(2) L_p 距离范数(L_p-norm),用于衡量扰动幅度[22,40],定义为

$$\|\delta\|_p=\left(\sum_{i=1}^{n}|\delta_i|^p\right)^{-p} \tag{11}$$

本文方法选择平均 L_2 范数作为整体扰动幅度的评估指标.扰动的 L_2 范数越小,意味着对初始图像修改的程度越小,更不容易被人类的视觉系统所感知.考虑到本文提出方法在黑盒攻击的过程中存在随机性,因此相关实验中都运行 5 次进行统计,并且记录测试数据的平均结果.

所有的黑盒攻击都在 Tensorflow 框架下进行测试.其中,

- 算法 1(ES-attack 方法)的参数设置如下:单位扰动幅值约束ε=0.063(图像像素值域归一化为 $x\in[0,1]$),最大迭代次数 $T=2\times10^5$,随机采样数量 M=10,协方差矩阵的标准差σ=0.005,步长α=1.0,协方差矩阵的维度 m=70×70×3,搜索路径学习率 c=0.01,协方差矩阵学习率 $c_1=c_2$=0.001.
- 算法 2(AM-Com 方法)的参数的设置如下:初始压缩率 r_0=0.75,最大压缩率 r_{max}=0.9,分组数量 N=100.

相关参数的详细讨论见第 3.2 节与第 3.3 节的实验.

3.2 ES-Attack方法相关参数实验

本节将验证 ES-Attack 黑盒攻击方法的有效性,并探讨不同的参数选择对结果的影响.经过实验发现,协方差矩阵的维度 m,搜索路径的学习率 c 与步长参数α的设置对攻击结果产生较大的影响.

- 协方差矩阵的维度 m 实验.

本文考虑 5 种大小的协方差矩阵维度 $m=l\times l\times3$,即 30×30×3,50×50×3, 70×70×3,90×90×3 与 110×110×3,通过 ES-Attack 攻击包括 Vgg-16、Inc-v3、Res-50 和 Res-101 在内的 4 种模型来研究维度参数 m 的影响.

表 1 记录了不同的协方差矩阵维度的攻击结果,包括平均查询次数 QN 与扰动的平均 L_2 范数,其中,搜索路径和协方差矩阵的学习率设置为 c=0.01 与 $c_1=c_2$=0.001.

由表 1 可以看出,4 种网络模型的攻击结果都呈现出相似的趋势:当 m 选择较低维度(即 l=30/50)的情况下,

ES-Attack 的攻击效率比高维度协方差矩阵(即 l=70/90/110)更低,需要交互查询的次数更多.例如在攻击 Inc-v3 网络模型的结果中,选择 l=30/50 时,ES-Attack 攻击平均需要交互查询次数为 QN=12452/7860;当选择 l=70/90/110 的情况下,ES-Attack 平均的交互查询次数为 QN=6941/6955/6961.同时,5 种矩阵维度下生成的扰动平均 L_2 范数分别为 L_2=15.27/14.62/14.73/14.91/14.97,扰动幅度差距相对较小.

Table 1 Results of different dimension parameter m (QN/L_2)

表 1 不同维度参数 m 的结果(QN/L_2)

Target model	l=30	l=50	l=70	l=90	l=110
Vgg-16	4 146/13.70	3 101/13.44	2 815/**13.37**	2 770/13.52	**2 759**/13.50
Inc-v3	12 452/15.27	7 860/**14.62**	**6 941**/14.73	6 955/14.91	6 961/14.97
Res-50	4 361/12.64	3 713/**12.38**	3 433/12.52	**3 406**/12.63	3 499/12.46
Res-101	4 874/12.67	4 294/**12.47**	**3 979**/12.60	4 150/12.67	4 087/13.03

经分析,主要由两种因素导致该结果.

(1) 当协方差矩阵维度 m 与输入图像数据的维度 n 差距过大时,低维度的协方差矩阵在更新搜索路径的过程中难以学习到足够精确的梯度方向分布信息.因此,协方差矩阵的维度大小与攻击效率呈正相关关系.

(2) 为了加速计算效率,ES-Attack 方法采取分层式攻击方法,先利用协方差矩阵进行随机采样得到 m 维样本 z,再经过线性插值得到 n 维偏移向量 u(算法 1 中第 5 行).上采样的过程会导致搜索路径方向信息的丢失,且维度越低的样本在上采样过程中丢失的信息越多,因此需要更多次的交互查询来重新搜索,导致了更低的攻击效率.

如公式(9)所示,协方差矩阵每次迭代更新消耗的计算资源为 $O(m^2)$,较低的维度 m 会减少资源的消耗,但增加交互查询次数.根据表 1 所示,当选择 l=70 时,协方差矩阵已经能学到较为精确的梯度分布信息,黑盒攻击的效率最优,且花费的计算资源相对更少.因此在本文实验中,选择以 m=70×70×3 作为最优协方差矩阵的维度,并作为后续实验的参数.

- 搜索路径的学习率 c 参数实验.

搜索路径学习了历史过程中采样向量的分布均值的移动方向,直接影响着黑盒攻击的效率.本文将对不同的学习率 c 进行讨论,考虑 4 种情况,即 c=0.005/0.01/0.03/0.05,其中,协方差的维度为 m=70×70×3,并将黑盒攻击 Vgg-16、Inc-v3、Res-50 和 Res-101 这 4 种模型的结果记录于表 2 中.

Table 2 Results of different learning rate c (QN/L_2)

表 2 不同学习率 c 的结果(QN/L_2)

Target model	c=0.005	c=0.01	c=0.03	c=0.05
Vgg-16	2 803/**13.30**	2 815/13.37	**2 758**/13.45	2 812/13.38
Inc-v3	7 212/14.97	**6 941**/**14.73**	7 107/14.88	7 234/14.76
Res-50	3 563/12.54	3 433/**12.52**	3 422/12.60	**3 369**/12.56
Res-101	4 246/**12.58**	**3 979**/12.60	4 063/12.71	3 986/12.67

由表 2 可得出结论:

(1) 学习率 c 的大小对生成扰动的平均 L_2 范数影响较少.例如在攻击 Res-101 模型的结果中,学习率 c=0.005/0.01/0.03/0.05 时,生成扰动的平均幅度分别为 L_2=12.58/12.60/12.71/12.67,差异较小.其他 3 种模型也展示出相同的趋势.

(2) 设置不同的学习率 c 会对 ES-Attack 的效率产生较大的影响.例如在攻击 Inc-v3 模型时,4 种学习率的平均交互查询次数为 QN=7212/6941/7107/7234,差距较大.

通过对比表 2 的数据,发现 c=0.01 时的结果相对最优,因此作为搜索路径最优的学习率.

- 移动步长α参数实验.

步长参数控制了对抗样本每次迭代的更新距离,对攻击的效率和生成扰动的幅度都有较大的影响.本文考虑了 5 种不同量级的步长,即α=0.01/0.1/1.0/10/100.表 3 记录了基于不同步长参数对 4 种模型 Vgg-16、Inc-v3、

Res-50 和 Res-101 的攻击结果.

Table 3 Results of step size α (QN/L_2)

表 3 不同步长α的结果(QN/L_2)

Target model	α=0.01	α=0.1	α=1.0	α=10	α=100
Vgg-16	25 575/**6.56**	9 710/8.77	2 815/13.37	**2 803**/14.88	2 945/14.94
Inc-v3	43 694/**7.95**	18 829/9.39	6 941/14.73	**6 732**/19.41	7 074/19.66
Res-50	25 781/**6.66**	10 844/7.95	**3 433**/12.52	3 978/14.76	3 661/14.79
Res-101	28 581/**7.06**	12 041/8.22	**3 979**/12.60	4 134/14.69	4 146/14.79

根据表 3 可以发现,步长参数对攻击结果的影响如下.

(1) 步长的量级与攻击的效率呈现出正相关的趋势:当α选择较小值(即α=0.01/0.1)的情况下,ES-Attack 的平均交互查询的次数较多,攻击效率较低;而当α大于 1 后,交互查询的次数收敛.

(2) 步长的量级会影响扰动的幅度:随着步长大小的增加,最终生成扰动的幅度也逐渐增大.当设置α=0.01 时,攻击预训练模型生成扰动的平均幅度最小,分别为 L_2=6.56/7.95/6.66/7.06.

本文综合考虑攻击效率与扰动幅度,选择α=1.0 作为后续实验的参数.

3.3 AM-Com方法相关参数实验

本节主要对 AM-Com 方法的有效性进行验证.AM-Com 首先基于替代网络模型生成的热力图对扰动进行分组,然后依次压缩各组的冗余扰动.其中,热力图的选择策略和分组数量 N 对压缩效果有较大的影响.

- 分组策略选择实验

本文选择了 4 种不同的热力图作为分组策略进行对比,分别为:

(1) 黑盒攻击的目标标签对应的热力图 $CAM(\tilde{y})$.

(2) 图像真实标签对应的热力图 $CAM(y)$.

(3) 目标标签与真实标签对应热力图的求和平均 $|CAM(y)+CAM(\tilde{y})|/2$,记为 $CAM(+)$.

(4) $|CAM(y)-CAM(\tilde{y})|$,即真实标签与目标标签热力图之差的绝对值,记为 $CAM(-)$.

此外,本文还考虑加入两种基线分组策略进行对比.

(5) 随机分组方法[26],即将扰动的每一维数据随机划分到任意一组,平均每组的数据数量为 n/N,记为 *Random*.

(6) Shi 等人[40]提出的 Whey 分组方法,即按照扰动向量的绝对值从小到大平均分为 N 组,记为 *Whey*.

图 4 展示了 Inc-v3 模型对应分组(1)~分组(4)策略下生成的热力图可视化示例,通过观察可以看出,$CAM(y)$ 主要关注的区域为图像内容的主体(如萨摩耶犬与金鱼的头部), $CAM(\tilde{y})$ 则主要关注图像主体之外的区域(如背景部分),而 $CAM(+)$与 $CAM(-)$的关注区域则介于二者之间.

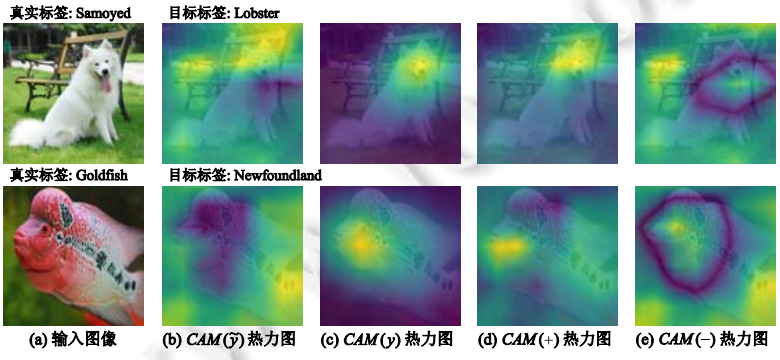


Fig.4 Visualized examples of different heatmaps generated by Inc-v3

图 4 Inc-v3 模型生成不同热力图的可视化示例

与第 3.2 节相同,用于生成热力图的替代网络模型包括 Vgg-16、Inc-v3、Res-50 和 Res-101,分组数量 N=100,

并以 ES-Attack 黑盒攻击 Res-50 模型生成的扰动进行压缩作为示例,结果记录在表 4 中,包括平均查询次数 *QN* 与优化后的平均扰动 L_2 范数.作为对比,未经过压缩的对抗样本的平均扰动 L_2 范数记录于"w/o *CAM*"栏.

Table 4 Results of compression with different grouping strategies (QN/L_2)

表 4 不同的分组策略压缩优化结果(QN/L_2)

f_s	w/o *CAM*	*CAM*($\tilde{y}$)	*CAM*(*y*)	*CAM*(+)	*CAM*(−)	*Random*	*Whey*
Vgg-16	−/12.52	287/11.64	282/**11.60**	**249**/11.63	261/11.65	285/11.70	286/11.66
Inc-v3		278/11.52	277/**11.51**	**237**/11.57	263/11.59		
Res-50		299/11.53	292/**11.39**	**250**/11.50	269/11.60		
Res-101		287/11.62	291/**11.51**	**247**/11.63	271/11.59		

由表 4 可知:

(1) 相比于 *Random* 和 *Whey* 这两种基线分组方法,本文结合注意力机制提出的 AM-Com 方法在 *CAM*($\tilde{y}$)、*CAM*(*y*)、*CAM*(+)和 *CAM*(−)这 4 种热力图分组策略下都取得了更优的压缩效果.例如,基于 Inc-v3 进行分组,相比未经优化的扰动幅度(L_2=12.52),压缩后减少的幅度分别为 L_2=1.0/1.01/0.95/0.93.而 *Random* 和 *Whey* 方法减少的扰动幅度相对较少,分别为 L_2=0.82/0.86,验证了 AM-Com 方法的有效性.

(2) *CAM*(*y*)和 *CAM*(+)这两种分组策略在所有替代模型的实验结果中分别表现出了最优压缩的效果/速度,即 *CAM*(*y*)策略减少的 L_2 幅度最多,以及 *CAM*(+)策略所需要的交互查询次数最少.相对的,分组策略 *CAM*(−)的压缩效果最差,与 *Whey* 策略的结果接近;而 *CAM*($\tilde{y}$) 策略花费的交互次数最多,因此优化效率最低.

(3) 尽管预训练模型之间的注意力区域存在差异,但基于不同的替代模型进行分组优化,都可以得到较优的压缩结果.其中,基于 Res-50 分组属于白盒优化,因此压缩效果最优;同时,根据图 3 的结果,Vgg-16 与 Res-50 的热力图相似度最低,因此黑盒优化的效果相对较差,进一步验证了 AM-Com 方法的鲁棒性和泛用性.

由于本文提出的方法在 AM-Com 阶段更注重压缩优化的效果,因此选择 *CAM*(*y*)作为最优分组策略,并用于后文的实验.

- 分组数量 *N* 选择实验

另一个对扰动压缩优化的结果产生较大影响的参数是分组数量 *N*.本节考虑 5 种分组数量进行实验,即 *N*=40/60/80/100/120.基于不同模型的 AM-Com 压缩结果如图 5 所示,其中,横轴代表交互查询次数 *QN*,纵轴表示优化后的扰动幅度 L_2 范数,各节点代表不同分组数量 *N* 对应的结果.

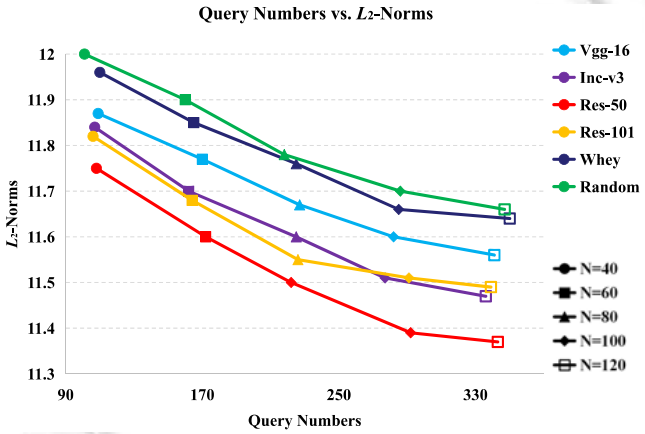


Fig.5 Results of compression with different group number *N* (QN/L_2)

图 5 不同分组数量 *N* 的压缩优化结果

从图 5 可以看出,所有的模型都表现出相同的趋势:随着分组数量 N 的增加,压缩过程中需要进行交互的查询次数也线性增加,并且最终优化后的扰动幅度更小.同时,在所有的分组数量下,基于不同模型进行 AM-Com 压缩优化的结果都比 *Whey* 和 *Random* 这两种基线策略效果更好.另一方面,与表 4 的结论类似,由于不同模型的热力图存在差别,导致黑盒优化会稍差于白盒优化的效果,其中,基于热力图相似性最低的 Vgg-16 模型进行优化,结果最差;反之,Res-50 模型(红色曲线)为白盒优化,因此各分组数量下都表现出最优的效果.该实验进一步验证了 AM-Com 的有效性.为了较好地权衡交互查询次数和扰动幅度,本文选择 N=100 作为最优的分组数量.

- AM-Com 方法的有效性分析

不同于 *Whey* 和 *Random* 策略,AM-Com 基于分类模型的热力图对扰动进行分组,考虑了注意力机制与冗余信息的内在联系,以此获取更好的压缩优化效果.本文将扰动的 L_2 密度(density of perturbation,简称 DoP)定义为区域内单位像素的扰动幅度,公式表示如下:

$$Dop(i)=\|\delta'[i]\|_2/|\delta'[i]| \tag{12}$$

其中,$\delta'[i]$是根据分组策略将扰动数据划分到第 i 组的集合.DoP 值反映了该区域内扰动的密集程度,该值越大,则说明扰动在该区域内越集中.

图 6 展示了 Vgg-16、Inc-v3、Res-50 和 Res-101 这 4 种模型对扰动结果进行分组后的平均 DoP 分布.其中,横轴为根据热力图数值分组的索引,索引越大,则该组区域的关注程度越高;纵轴为各组对应的 DoP 值,绿色和红色区域分别为使用 AM-Com 方法进行压缩前后的扰动密度.根据图 6 分析,可以得出以下结论.

(1) ES-Attack 方法生成的扰动(绿色区域)在关注度最低和最高的区域都呈现出高密度趋势,可以推测出,修改图像对应区域的数据会对模型的预测结果产生较大影响.

(2) 经过 AM-Com 方法压缩优化后的扰动(红色区域)在关注度最低和最高的区域的 DoP 值显著降低,该结果证明了这些区域中存在着更多的冗余扰动;同时,观察图 1 和图 2 的可视化结果可以得出相同的结论,验证了 AM-Com 压缩方法的有效性.

(3) 优化后的扰动在高关注度区域的 DoP 值依然高于其他区域;反之,低关注区域的 DoP 值已经低于绝大多数区域,进一步证明了注意力机制与对抗样本扰动密切相关,即,对关高注度区域的数据添加扰动会产生更大的影响.

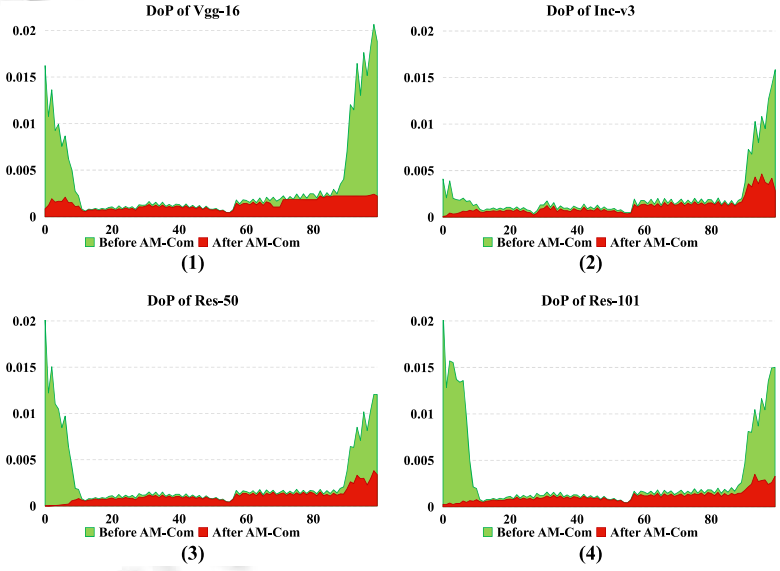


Fig.6 Distribution of average DoP based on four different models

图 6 4 种不同模型的平均 DoP 分布

3.4 相关工作对比实验

为了验证本文提出方法的高效性,将本文提出的方法与目前表现最优的 4 种黑盒攻击方法进行比较,包括:(1) 基于自动编码器的零阶优化方法 AutoZOOM[28];(2) 基于有限差分算法的对抗攻击方法 FD-Attack[26];(3) 基于自然进化算法的有限查询方法QL-Attack[29];(4) 基于模型决策边缘的攻击方法D-based Attack[33].为了进行公平的对比,本文需要对上述 4 种方法的参数进行设置.在 AutoZOOM 方法中,本文选择攻击效果最好的AutoZOOM-AE 策略进行实验[28],正则权重初始化为λ=10.0,且随着迭代次数增加.考虑到 FD-Attack 方法在对抗攻击大尺度图片时需要大量的查询次数,因此采用随机分组[26]策略进行加速,其中,FD-Attack 方法的随机分组的像素数量设为 1 000.QL-Attack 方法中,按原文实验设置,随机采样向量的数量 M=50,步长α=0.01,标准差σ=0.001.上述 3 种方法以首次目标攻击成功或达到最大查询次数 $QN_{\max}=2\times10^5$ 为停止条件.D-based Attack 方法的初始对抗样本图像选自 ImageNet 数据集,并可以被正确分类为目标标签.本文与 D-based Attack 方法对比两种结果:(1) 限制查询次数与本文方法的结果相同时,生成扰动的 L_2 范数,记为 D-based;(2) 当查询次数达到上限 $QN_{\max}$ 时,生成扰动的平均 L_2 范数,记为 D-based+.

黑盒攻击的目标模型包括 4 种预训练模型:Vgg-16、Inc-v3、Res-50、Res-101 和 3 种对抗防御模型:Adv-Inc、Inc-v3$_{ens3}$、Inc-v3$_{ens4}$.为了验证提出方法的泛化性,在 AM-Com 压缩方法中,本文分别基于 Vgg-16、Inc-v3、Res-50 和 Res-101 这 4 种模型生成热力图压缩优化对抗样本的扰动.对比的数据包括平均查询次数 QN 与优化后的扰动 L_2 范数,记录在表 5 中.其中,每一行表示不同方法对同一个网络模型攻击的结果,最优的数据以粗体表示.本文方法结果的对角线数据为白盒优化(例如以 Inc-v3 模型为基础来优化 Inc-v3 模型生成的对抗样本),考虑到黑盒场景中的模型不可知性约束,因此不加入对比实验,以“–”符号表示.

Table 5 Comparison with related methods (QN/L_2)

表 5 与相关方法的对比结果(QN/L_2)

目标模型	本文方法				相关方法				
	Vgg-16	Inc-v3	Res-50	Res-101	**Auto ZOOM**	**FD attack**	**QL attack**	**D-based**	**D-based+**
Vgg-16	–	**3 078** 12.34	3 091 12.19	3 090 12.26	7 420 22.12	68 508 **8.95**	7 680 15.95	3 090 32.06	– 21.61
Inc-v3	**7 233** 14.01	–	7 238 13.97	7 242 13.96	13 068 34.38	128 457 **10.52**	11 160 20.33	7 242 48.79	– 29.49
Res-50	3 715 11.60	**3 710** 11.51	–	3 724 11.51	8 087 23.89	56 258 **8.63**	6 990 15.77	3 724 38.03	– 22.90
Res-101	4 283 11.73	**4 279** 11.68	4 292 11.59	–	8 439 25.21	61 059 **8.78**	7 620 15.70	4 292 36.35	– 23.13
Adv-Inc	**12 692** 16.06	12 695 16.04	12 698 16.07	12 703 **16.03**	17 023 41.39	\	36 110 20.22	12 703 72.49	– 58.57
Inc-v3$_{ens3}$	10 111 15.62	10 115 **15.58**	**10 110** 15.61	10 113 15.63	16 864 42.28	\	21 610 20.37	10 115 64.11	– 49.68
Inc-v3$_{ens4}$	**9 320** 15.67	9 335 **15.56**	9 340 15.57	9 338 15.59	15 350 39.50	\	16 767 20.08	9 320 60.62	– 48.39

由表 5 可以看出,在 4 种相关的黑箱攻击方法中,

- FD-Attack 需要花费大量的查询次数来换取较小的扰动幅度,攻击效率最低;且在攻击对抗防御模型时会超出最大查询次数,导致无法攻击成功(表 5 中 FD-attack 列以“＼”符号表示),难以实际应用.
- AutoZOOM 方法的攻击效率较优,特别是攻击对抗训练的防御模型时,交互查询次数少于另外 3 种方法.但平均扰动幅度较大,容易被人眼察觉,如图 7 所示.
- QL-Attack 方法黑箱攻击的效率较高,且生成的扰动幅度也相对较小,综合表现较优.
- D-based Attack 方法与 FD-Attack 类似,需要花费大量的查询次数来逐步减小扰动幅度;且当查询次数达到最大限制时,扰动幅度依然较大.此外,该方法在攻击防御模型时降低的扰动幅度很小,难以满足不可感知约束,隐蔽性较差.

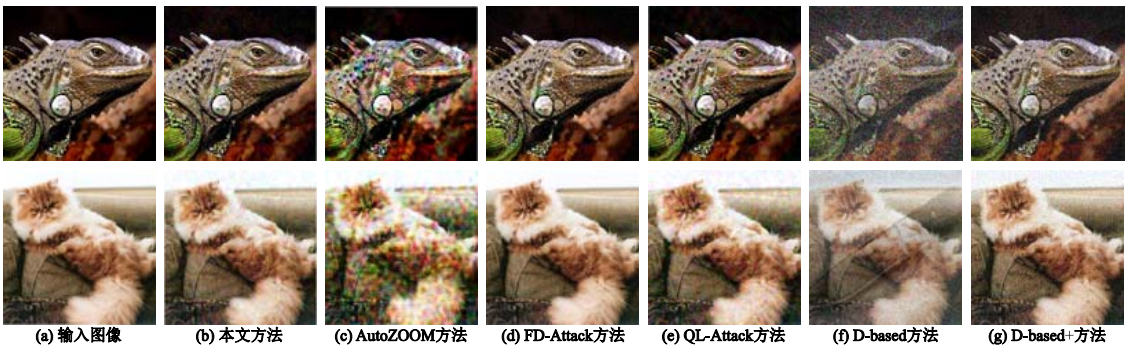


Fig.7 Adversarial examples generated from different methods for attacking Inc-v3 model

图 7 不同方法攻击 Inc-v3 模型生成的对抗样本

本文提出的方法在黑盒场景下对包括对抗防御模型在内的 7 种分类器都可以完成有效的攻击,且基于不同模型的热力图对扰动进行分组压缩都表现出较好的优化结果.通过与相关方法进行对比,可以发现本文方法所需要的平均查询次数与扰动幅度都展现出较优的效果.

- 在攻击 4 种预训练模型的结果中,相比于表现最优的 QL-Attack 方法,本文方法分别减少了 60.0%、35.2%、46.9%、43.9%的平均查询次数;相比于效率较低的 FD-Attack 方法,更是减少了 95.5%、94.6%、93.4%、93.0%的平均查询次数,攻击效率相对提升了 10 倍以上.
- 在攻击 3 种对抗防御模型的结果中,本文方法比表现最佳的 AutoZOOM 方法分别减少了 25.5%、40.0%、39.3%的平均查询次数,因此在攻击效率上有了较大的提升.
- 同时,基于不同模型优化后的平均扰动幅度也都小于 QL-Attack 方法和 AutoZOOM 方法.
- 在相同的查询次数限制下,本文方法生成扰动的幅度也远小于 D-based Attack 方法(详见表 5 中 D-based 列);冗余的扰动信息更少,仅次于扰动幅度最小的 FD-Attack 方法.
- 如图 7 所示,从视觉上观察,本文提出方法产生的对抗样本噪点不明显,比 AutoZOOM 方法、QL-Attack 方法和 D-based Attack 方法的结果更接近原始图像,降低了人眼视觉察觉的可能性.

该实验验证了本文提出方法的有效性和鲁棒性,相比于目前最优的 4 种黑盒攻击方法,在平均查询次数与扰动幅度指标上都有明显的提升.

3.5 对抗训练实验

本节主要验证本文方法在对抗训练中提升模型鲁棒性的效果.考虑到从零训练高维度图像数据的对抗防御模型需要消耗高额的计算资源[42],本文结合 Goodfellow 等人[19]的混合训练策略与 Xie 等人[43]的方法对现有的基于 PGD[44]训练的防御模型(Adv-Inc、Inc-v3$_{ens3}$、Inc-v3$_{ens4}$)进行对抗微调.训练过程主要包含 3 个步骤:首先从 ImageNet 验证集中随机选取测试数据以外的 10 000 张图像作为原始样本,再利用 ES-Attack 方法攻击预训练模型生成相应的对抗样本,最后用对抗样本形成新的数据集(不包含原始样本)对已有的鲁棒模型进行 5 个 epoch 微调,得到新的对抗模型 Adv-Inc*、Inc-v3$_{ens3}^*$和 Inc-v3$_{ens4}^*$.

为验证对抗微调后防御模型的鲁棒性,本文选择第 3.4 节实验中表现较优的 3 种黑盒攻击方法(ES-attack、AutoZOOM、QL-attack)对 Adv-Inc*、Inc-v3$_{ens3}^*$和 Inc-v3$_{ens4}^*$模型进行攻击,相关参数设置相同,并将平均查询次数 QN 与优化后的平均扰动 L_2 范数记录于表 6 中.

对比表 5 与表 6 的数据可知,经过 ES-Attack 对抗训练后模型的鲁棒性得到了进一步提升.相比于微调前的对抗防御模型,上述 3 种方法在攻击 Adv-Inc*、Inc-v3$_{ens3}^*$和 Inc-v3$_{ens4}^*$模型时都需要花费更多的计算资源,例如 AutoZOOM 方法分别增加了 49.0%、33.0%、38.4%的平均查询次数,而且生成的对抗样本扰动幅度也相对更大.该实验同时展示了一个有趣的现象,即仅利用一种黑盒攻击(ES-attack)对网络模型进行对抗训练,也可以提高

该防御模型抵抗其他黑盒攻击(AutoZOOM、QL-attack)的能力.

Table 6 Results of fine-tuned adversarial defense models (QN/L_2)

表 6 微调后的对抗防御模型结果(QN/L_2)

目标模型	本文方法				相关方法	
	Vgg-16	Inc-v3	Res-50	Res-101	**AutoZOOM**	**QL-attack**
Adv-Inc*	17 440 17.46	17 434 **17.37**	**17 431** 17.42	17 435 17.42	25 377 46.39	52 439 21.11
Inc-v3$_{ens3}$*	**14 745** 16.53	14 755 16.47	14 759 **16.44**	14 750 16.49	22 440 45.62	31 252 20.52
Inc-v3$_{ens4}$*	**13 630** 16.62	13 639 **16.55**	13 634 16.57	13 632 16.58	21 247 43.91	24 980 20.79

4 总 结

深度神经网络容易被对抗样本恶意攻击,而黑盒对抗攻击有着模型不可知和查询限制的约束,更符合实际的攻击场景,威胁着各类系统的安全性.目前,大多数的黑盒攻击方法存在着需要大量交互查询和生成扰动冗余过多的缺陷,导致攻击效率较低,且易于被人眼察觉,难以实际应用.针对于此,本文提出了一种基于协方差矩阵自适应进化策略的攻击方法,充分考虑了在攻击过程中梯度更新方向的内在联系,在迭代攻击过程中学习历史信息中较优的搜索路径,结合协方差矩阵,以高几率采样得到有效的扰动向量,降低交互查询的次数,提升攻击效率.在保持黑盒攻击成功的前提下,本文利用替代模型的类间激活热力图对生成的扰动分组,并依次进行压缩优化,减少在迭代攻击过程中积累的冗余扰动.本文结合实验分析了注意力机制与对抗样本的内在关联,证明了提出方法的可靠性.同时,本文方法还与 4 种表现最优的黑盒对抗攻击方法在 7 种深度神经网络模型上进行对比,实验结果展示出本文方法在平均查询次数与扰动幅度范数指标上都有较大的提升,充分验证了本文提出方法的有效性与鲁棒性.

References:

[1] Niu L, Veeraraghavan A, Sabharwal A. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2018. 7171−7180.

[2] Huang JP, Shi YH, Gao Y. Multi-scale Faster-RCNN algorithm for small object detection. Journal of Computer Research and Development, 2019,56(2):319−327 (in Chinese with English abstract).

[3] Huang L,Yang Y, Wang QJ, Guo F, Gao Y. Indoor scene segmentation based on fully convolutional neural networks. Journal of Image and Graphics, 2019,24(1):64−72 (in Chinese with English abstract).

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2014,39(4):640−651.

[5] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012,25:1097−1105.

[6] Simonyan K, Zisserman A, Very deep convolutional networks for large-scale image recognition. In: Proc. of the Int'l Conf. on Learning Representations. 2015. 1−14.

[7] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 2818−2826.

[8] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 770−778.

[9] Song M, Zhong K, Zhang J, *et al*. In-situ AI: Towards autonomous and incremental deep learning for IoT systems. In: Proc. of the Int'l Symp. on High Performance Computer Architecture (HPCA). IEEE, 2018. 92−103.

[10] Wang Y, Huang XD, Guo ST. Indoor fingerprint location algorithm based on convolutional neural network. Ruan Jian Xue Bao/ Journal of Software, 2018,29:63−72 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18007.htm

[11] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2014. 1−10.

[12] Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542(7639):115−118.

[13] Ma YK, Wu LF, Jian M, Liu FH, Yang Z. Algorithm 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]

[14] Wang WQ, Wang R, Wang LN, Tang BX. Adversarial examples generation approach for tendency classification on Chinese texts. Ruan Jian Xue Bao/Journal of Software, 2019,30(8):2415−2427 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5765.htm [doi: 10.13328/j.cnki.jos.005765]

[15] Sharif M, Bhagavatula S, Bauer L, *et al*. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2016. 1528−1540.

[16] Athalye A, Engstrom L, Ilyas A, *et al*. Synthesizing robust adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning. 2018. 284−293.

[17] Eykholt K, Evtimov I, Fernandes E, *et al*. Robust physical-world attacks on deep learning visual classification. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2018. 1625−1634.

[18] Thys S, Van Ranst W, Goedeme T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In: Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019. 1−7.

[19] Goodfellow IJ, Shlens J, Szegedy S. Explaining and harnessing adversarial examples. In: Proc. of the Computer Science. 2014. 1−11.

[20] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Proc. of the Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018. 99−112.

[21] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik Z, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy. IEEE, 2016. 372−387.

[22] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE, 2017. 39−57.

[23] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z, Swami A. Practical black-box attacks against machine learning. In: Proc. of the ACM Asia Conf. on Computer and Communications Security. ACM, 2017. 506−519.

[24] Dong Y, Pang T, Su H, *et al*. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2019. 4312−4321.

[25] Zhou W, Hou X, Chen Y, *et al*. Transferable adversarial perturbations. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 452−467.

[26] Bhagoji AN, He W, Li B, *et al*. Practical black-box attacks on deep neural networks using efficient query mechanisms. In: Proc. of the European Conf. on Computer Vision. Cham: Springer-Verlag, 2018. 158−174.

[27] Chen PY, Zhang H, Sharma Y, *et al*. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017. 15−26.

[28] Tu CC, Ting P, Chen PY, *et al*. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 742−749.

[29] Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information. In Proc. of the 35th Int'l Conf. on Machine Learning. 2018. 2137−2146.

[30] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans. on Evolutionary Computation, 2019, 23(5):828−841.

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

[32] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, *et al*. Universal adversarial perturbations. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. 2017. 1765−1773.

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

[34] Dong Y, Su H, Wu B, *et al*. Efficient decision-based black-box adversarial attacks on face recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2019. 7714−7722.

[35] Zhou B, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 2921−2929.

[36] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the IEEE Int'l Conf. on Computer Vision, Vol.7. 2017. 618−626.

[37] Liu A, Liu X, Fan J, *et al*. Perceptual-sensitive GAN for generating adversarial patches. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 1028−1035.

[38] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 2001,9(2): 159−195.

[39] Dong Y, Liao F, Pang T, *et al*. Boosting adversarial attacks with momentum. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2018. 9185−9193.

[40] Shi Y, Wang S, Han Y. Curls & Whey: Boosting black-box adversarial attacks. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2019. 6519−6527.

[41] Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE, 2009. 248−255.

[42] Florian T, Alexey K, Nicolas P, Dan B, Patrick M. Ensemble adversarial training: Attacks and defenses. In: Proc. of the Int'l Conf. on Learning Representations. 2018. 1−12.

[43] Xie C, Yuille A. Intriguing properties of adversarial training at scale. In: Proc. of the Int'l Conf. on Learning Representations. 2020. 1−14.

[44] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the Int'l Conf. on Learning Representations. 2018. 1−28.

附中文参考文献:

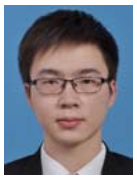
[2] 黄继鹏,史颖欢,高阳.面向小目标的多尺度 Faster-RCNN 检测算法.计算机研究与发展,2019,56(2):319−327.

[3] 黄龙,杨媛,王庆军,郭飞,高勇.结合全卷积神经网络的室内场景分割.中国图像图形学报,2019,24(1):64−72.

[10] 王英,黄旭东,郭松涛.基于卷积神经网络的室内指纹定位算法.软件学报,2018,29:63−72. http://www.jos.org.cn/1000-9825/18007.htm

[13] 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]

[14] 王文琦,汪润,王丽娜,唐奔宵.面向中文文本倾向性分类的对抗样本生成方法.软件学报,2019,30(8):2415−2427. http://www.jos.org.cn/1000-9825/5765.htm [doi: 10.13328/j.cnki.jos.005765]



黄立峰(1990－),男,博士生,CCF 学生会员,主要研究领域为对抗学习,自主感知定位.

廖泳贤(1996－),女,硕士生,主要研究领域为对抗训练,计算机视觉.

庄文梓(1997－),男,硕士生,主要研究领域为对抗训练,计算机视觉.

刘宁(1973－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为对抗学习,自主感知定位.