



计算机工程  
Computer Engineering  
ISSN 1000-3428, CN 31-1289/TP

## 《计算机工程》网络首发论文

题目: 基于模型间迁移性的黑盒对抗攻击起点提升方法  
作者: 陈晓楠, 胡建敏, 张本俊, 陈爱玲  
DOI: 10.19678/j.issn.1000-3428.0059105  
网络首发日期: 2020-10-14  
引用格式: 陈晓楠, 胡建敏, 张本俊, 陈爱玲. 基于模型间迁移性的黑盒对抗攻击起点提升方法. 计算机工程. <https://doi.org/10.19678/j.issn.1000-3428.0059105>



**网络首发:** 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式 (包括网络呈现版式) 排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

**出版确认:** 纸质期刊编辑部通过与《中国学术期刊 (光盘版)》电子杂志社有限公司签约, 在《中国学术期刊 (网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊 (网络版)》是国家新闻出版广电总局批准的网络连续型出版物 (ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。



# 基于模型间迁移性的黑盒对抗攻击起点提升方法

陈晓楠<sup>1</sup>, 胡建敏<sup>1\*</sup>, 张本俊<sup>2</sup>, 陈爱玲<sup>3</sup>


(1. 国防大学 联合勤务学院, 北京 100089;

2. 南京航空航天大学 电子信息工程学院, 江苏南京 211100;

3. 山东省烟台市实验中学, 山东烟台 265500)

**摘 要:** 对抗样本的发现, 给神经网络的安全性带来了新的挑战。为了高效地寻找基于决策的黑盒攻击下的对抗样本, 提出了一种利用模型之间的迁移性来提升对抗起点的方法。通过模型之间的迁移性来循环叠加干扰图像, 生成初始样本作为新的攻击起点来进行边界攻击, 实现了基于决策的无目标黑盒对抗攻击和有目标黑盒对抗攻击。实验结果表明, 无目标攻击节省了 23% 的查询次数, 有目标攻击节省了 17% 的查询次数, 并且整个黑盒攻击算法所需时间低于原边界攻击算法的所耗费的时间。

**关键词:** 黑盒攻击; 对抗样本; 迁移性; 初始样本; 边界攻击; 无目标攻击; 有目标攻击

开放科学(资源服务)标志码(OSID): 

## Black box attack starting point promotion method based on mobility between models

CHEN Xiaonan<sup>1</sup>, HU Jianmin<sup>1\*</sup>, ZHANG Benjun<sup>2</sup>, CHEN Ailing<sup>3</sup>

(1. Joint Logistics College, National Defense University, Beijing 100089, China;

2. College of Electronic Information Engineering, Nanjing University of Aeronautics and Astronautics, Jiangsu Nanjing 211100, China;

3. Yantai Experimental Middle School of Shandong Province, Shandong Yantai, 265500, China)

**【Abstract】** The discovery of counter samples brings new challenges to the security of neural networks. In order to find the confrontation samples under the black box attack based on decision efficiently, a method is proposed to enhance the confrontation starting point by using the mobility between models. The initial samples are generated as a new attack starting point to carry out boundary attack. The decision-making based no target black box counter attack and targeted black box counter attack are realized. The experimental results show that the no target attack saves 23% of the query times, and the targeted attack saves 17% of the query times, and the whole black box attack algorithm takes less time than the original boundary attack algorithm.

**【Key words】** black box attack; counter sample; mobility; initial sample; boundary attack; no target attack; targeted attack

DOI:10.19678/j.issn.1000-3428.0059105

### 1 概述

人工智能作为当今最热门的领域之一, 在很多领域呈现爆发式的发展趋势, 深度学习已经成为目前应用最广泛的技术之一, 在数据挖掘、自然语言处理、计算机视觉、信息检索等领域中占有很重要的地位, 但是很多研究表明, 神经网络容易被精心设计的微小扰动所影响, 通过在正常样本上叠加微小恶意噪声而成为的对抗样本进行对抗攻击, 导致模型输入错误的结果, 进而影响到

实际应用系统的安全性。对抗攻击和对抗样本的发现使人工智能领域面临着巨大的安全威胁。

随着对深度学习的深入研究, 大量的对抗攻击方法相继被提出, 对抗攻击主要可以分为白盒攻击和黑盒攻击两类:

1. 白盒攻击: 攻击者完全掌握目标模型的内部结构和训练参数值, 甚至还包括其特征集合、训练方法和训练数据等来进行的对抗攻击。

**基金项目:** 中央大学基础研究基金 (NP2019102), 中国航空科学基金会 (201943052001)

**作者简介:** 陈晓楠(1991—), 男, 山东烟台人, 硕士研究生, 主要研究方向: 联合勤务管理、智能后勤; 胡建敏(1964—), 男, 河北吴桥人, 博士生导师, 教授, 主要研究方向: 军事管理学; 张本俊(1996—), 女, 安徽六安人, 硕士研究生, 主要研究方向: 信息与通信工程; 陈爱玲(1978—), 女, 山东烟台人, 本科, 主要研究方向: AI 教育。E-mail: cxn6300296@163.com

2.黑盒攻击:攻击者在不知道目标模型的内部结构、训练参数和相应算法的情况下,通过数据的输入与输出结果来进行分析,设计和构造对抗样本,实施对抗攻击。在实际的运用场景中,黑盒攻击的安全威胁更为严峻。

深度学习模型能够以达到甚至高于人眼的识别度来识别图像,但是对抗攻击和对抗样本的发现,将导致模型识别错误。这些漏洞如果被别有用心的人所掌握,将可能产生严重的安全问题,例如,自动驾驶汽车的识别错误,将可能导致严重的交通事故<sup>[1]</sup>,某些犯罪分子可能利用生成的对抗样本逃过人脸识别的检测等<sup>[2,3]</sup>。

在文献[4]中, Szegedy C 等人证明了将图片进行适当修改后能够使深度学习模型识别错误;在文献[5]中, Goodfellow I J 提出了产生对抗攻击根本原因的猜测——深度神经网络在高维空间中的线性特性已经足以产生这种攻击行为,并提出了快速梯度符号方法 (FGSM), 作为经典的攻击手段;在文献[6]中, Minhao Cheng 等人提出了 SIGN -OPT 算法,使得黑盒攻击对目标模型的查询次数大大降低;在文献[7]中, Jianbo Chen 等人提出了 HopSkipJumpAttack2 算法具有很高的查询效率;在文献[8]中, Dong Y 等人在 FGSM 算法基础上,提出基于动量的迭代算法 T-MI-FGSM 提高对抗样本的可迁移性和黑盒攻击的成功率;在文献[9]中, Sarkar S 提出了 UPSET 和 ANGRI 算法,分别使用残差梯度网络构造对抗样本和生成图像特定的扰动,产生了很高的欺骗率;在文献[10]中, Akhtar N 等人对计算机视觉中的对抗攻击进行了较为详尽的研究,提出了 12 种对抗攻击的方法和 15 种防御的措施;在文献[11]中,分析并汇总了近年以来关于深度学习对抗攻击的部分算法并进行了比较与分析。总而言之,关于生成对抗样本进行对抗攻击的研究很多,这些研究的成果促进了深度学习的进一步发展。

文献[12]首次提出了基于决策的边界攻击方法,且该方法适合绝大多数的神经网络模型,具有很好地通用性和普适性,文献[7]在基于决策基础上对决策边界处的梯度方向进行估计,并提出了控制偏离边界的误差的方法,有效提高基于决策的攻击效率,文献[13]提出了一种新的基于决策的攻击算法,它可以使用少量的查询生成对抗性示例。无论是文献[7]、文献[12]还是文献[13],都是逐步朝着更优解靠拢,本质上是一样的,关键在于靠拢的方式不同,本文同样是基于决策算法的基础上实现无目标攻击和有目标攻击,与这几个文献不同的是,本文提出可以利用模型的迁移性来循环叠加干扰图像,找到新的初始样本,提高了文献[7]、文献[12]还是文献[13]中的基于决策算法的运算起点,大大降低了查询次数。

## 2 基于决策的黑盒攻击思路

一般来讲,对抗攻击可以分为四大类:基于梯度的攻击、基于分数的攻击、基于迁移的攻击和基于决策的攻击。在前两攻击中,基于梯度的攻击多用于白盒攻击中,基于分数的攻击多用于黑盒攻击中,在很多黑盒攻击中,攻击者可以通过对目标模型的输入来观察分类结果,并可以获得每个类别的分数,攻击者可以根据这个分数来设计对抗样本的生成算法。

文献[14]和文献[15]中指出对抗样本具有迁移性:相同的对抗样本,可以被不同的分类器错误分类,即基于迁移的攻击。一般来讲,基于迁移的攻击是利用训练数据训练一个可以替代的模型,即对抗样本在不同的模型之间可以进行迁移。

与其他几种攻击相比,基于决策的攻击与实际应用更为相关。与此同时,基于决策的攻击比其他攻击类型更高效、更稳健、更难以被察觉。在文献[12]中,引入了一个普适性的对抗攻击算法:边界攻击。文献中提到的边界攻击是属于基于决策攻击的一种,适合对复杂自然数据集的有效攻击,其原理是沿着对抗样本和正常样本之间的决策边界,采用比较简单的拒绝抽样算法,结合简单的建议分布和信赖域方法启发的动态步长调整。边界攻击的核心是从一个较大的干扰出发,逐步降低干扰的程度,这种思路基本上推翻了以往所有的对抗攻击的思路。

与其他三种攻击相比,基于决策的攻击只需要模型的输入输出类别,并且应用起来要简单得多。本文就是基于决策攻击来设计的算法,仅仅知道分类的结果,不能得到每个类别的分数,以此来设计黑盒算法生成对抗样本、进行对抗攻击。

这种情况下,黑盒攻击的一般思路是先使用现成的模型去标记我们的训练数据,选择一副图片,输入到模型中,通过模型反馈的标签来当作监督信号,不断地变换图片,形成一个新的数据集,图片的选择可以用真实的图片,也可以用合成的图片,将这个数据集作为训练集,在这个训练集上训练出一个新的模型,这个新的模型对我们来说是透明的,在这个模型上采用白盒攻击的手段生成对抗样本,这个样本也有极高的概率能够骗过原先的模型。

当然,这个方法可以更简单一些。选择一幅图像,然后给这个图像加一些噪声,通过不断地变换噪声,使得模型输出的分类结果发生改变,实际上此时我们的图片已经碰到了模型的分界边界。接下来就是不断地尝试,找到一个能让分类结果错误的最小的噪声



上述两种情况都有一个共同的问题,就是需要大量的向目标模型进行查询,查询到可以构建自己的训练集的程度,为了能够尽可能的减少查询,本文对传统的边界攻击加以改进与完善,提出了一种通过模型之间的迁移性来循环叠加干扰图像来确定初始样本,然后采用边界攻击生成对抗样本的算法,目的是为了提高传统边界算法的运算起点,尽可能的减少查询数量,更好地欺骗原模型的分能力。

已知一个神经网络模型,我们对其内部结构和参数一无所知,唯一知道模型的作用是进行图片分类的,并且不知道这个模型的分能力。不知道这个模型是一个最简单的神经网络,只能够进行简单的分类任务,还是一个复杂的 DNN、CNN 或者 RNN 神经网络,可以完成非常复杂的分类任务。

本文提出的改进黑盒算法当中最重要的是初始样本的确定。文献[7]、文献[12]、文献[13]中,都是逐步朝着更优解靠拢,本质上是一样的,关键在于靠拢的方式不同,无目标攻击的初始样本是一个随机扰动图片,目标攻击的初始样本是一个指定的目标分类的图片,如图 1 和图 2 所示。

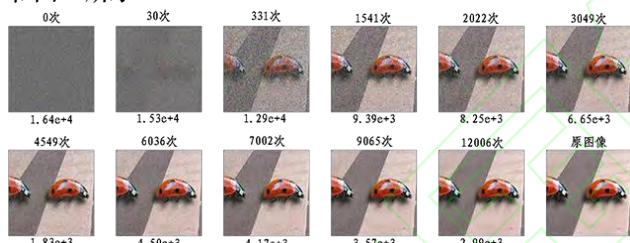


图 1 无目标攻击过程

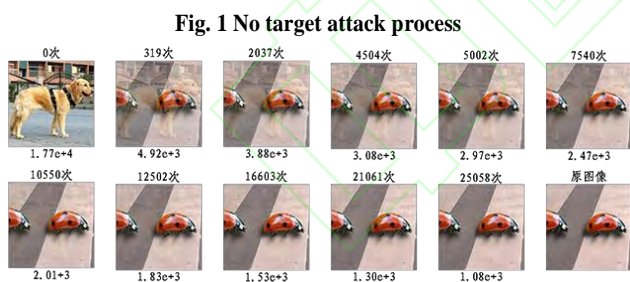


图 2 有目标攻击过程

Fig. 2 Target attack process

图的上方是对目标神经网络的查询次数,下方是与原图像之间的均方误差。可以看出,在完全黑盒状态下,如果要完善的构造出有效的对抗样本,则需要查询数千次甚至上万次才能逐步找到一张足够清晰的图片来骗过神经网络。文献[7]、文献[12]、文献[13]还有类似的大多数文献都是针对这个逐步查询的过程来进行优化的,从而更快的靠近目标样本,使得整体的查询次数降低。如果可以找到一个更高无目标攻击和有目标攻击的运算起

点,不从随机干扰图像和指定类别的额图像出发,将大大降低对目标神经网络的查询次数。接下来就是确定新的初始样本来提升整体攻击的起点,进而降低整体的查询次数。

### 3 基于决策的黑盒攻击原理与流程

#### 3.1 算法基础

基于决策的黑盒攻击具体分为两步,第一步是通过模型之间的迁移性来多次叠加干扰图像的方法确定初始样本,第二步在初始样本的基础上运用边界攻击的手段来确定合适的对抗样本。这里的通过模型之间的迁移性来进行的攻击并不能算是真正的迁移攻击,迁移攻击利用的是训练数据的信息,而这里采用的是根据部分查询结果来自己收集、扩充和构造数据集,进行对应的二分类训练,分类项为目标图片类和非目标图片类,然后采用 FGSM、I-FGSM、MI-FGSM、DI<sup>2</sup>-FGSM、M-DI<sup>2</sup>-FGSM 等算法来生成对抗样本进行叠加尝试,进而确定合适的初始样本。

快速梯度符号方法 (FGSM) 是非常经典的一种攻击方法<sup>[5]</sup>,通过计算交叉熵损失的梯度  $J(x, y_T)$  来找出噪声的方向

$$x' = x + \varepsilon \cdot \text{sign}(\nabla J(x, y_T)). \quad (1)$$

I-FGSM 将噪声  $\varepsilon$  的上限分成几个小的步长  $\alpha$ , 并逐步增加噪声<sup>[16]</sup>

$$x'_{t+1} = \text{Clip}_{x, \varepsilon} \{x'_t + \alpha \cdot \text{sign}(\nabla J(x'_t, y_T))\}. \quad (2)$$

I-FGSM 在白盒场景中的所有当前迭代攻击中具有最高的攻击效果。其主要缺点是迭代步骤的边际效应递减。具体是随着迭代次数  $t$  的增加和步长  $\alpha$  的减小,保持加入迭代步骤对攻击效果的改善很小。

DI<sup>2</sup>-FGSM 在 I-FGSM 基础上,将动量集成到攻击过程中,稳定更新方向,在一定程度上避免了局部极值

$$x'_{t+1} = \text{Clip}_{x, \varepsilon} \{x'_t + \alpha \cdot \text{sign}(\nabla J(T(x'_t; p), y_T))\}. \quad (3)$$

MI-FGSM 引入了一个动量项,使噪声添加方向的调整更加平滑,但边际效应递减对迭代次数的影响仍然存在<sup>[8]</sup>

$$m_{t+1} = \mu \cdot m_t + \frac{\nabla J(x'_t, y_T)}{\|\nabla J(x'_t, y_T)\|}, \quad (4)$$

$$x'_{t+1} = \text{Clip}_{x, \varepsilon} \{x'_t + \alpha \cdot \text{sign}(\nabla J(g_{t+1}))\}.$$

M-DI<sup>2</sup>-FGSM 在 MI-FGSM 基础上,将动量集成到攻击过程中,稳定更新方向,在一定程度上避免

了局部极值<sup>[17]</sup>

$$m_{t+1} = \mu \cdot m_t + \frac{\nabla J(T(x'_t; p), y_T)}{\|\nabla J(T(x'_t; p), y_T)\|}, \quad (5)$$

$$x'_{t+1} = \text{Clip}_{x, \varepsilon} \left\{ x'_t + \alpha \cdot \text{sign}(\nabla J(g_{t+1})) \right\}.$$

以上的 FGSM 系列算法属于白盒攻击算法, 需要掌握模型的内部结构和训练参数值, 针对的是自己搭建的二分类神经网络, 目的是为了获得一个可以接纳的初始样本, 提升黑盒攻击的攻击起点, 具体算法流程如下所示。

#### 算法 1 初始样本叠加算法

##### Algorithm 1. Initial sample superposition algorithm

**算法:** 初始样本叠加算法

**输入:** 目标图像  $o$ , 二分类神经网络

**输出:** 对抗样本  $\tilde{o}$

**初始值:**  $\tilde{o}^0 = o$ ,  $k = 0$ ,  $\tilde{o}^k$  为第  $k$  次生成的对抗样本  
通过 FGSM 系列算法生成对抗样本  $\tilde{o}^1 = \tilde{o}^0 + \eta_0$

**While**  $\tilde{o}^{k+1}$  和  $\tilde{o}$  不是对抗关系

将  $\tilde{o}^{k+1}$  进行旋转、镜像、缩放等数据增强技术进行数据扩充,  
然后投入到数据集中

通过增量训练更新二分类神经网络并导出权重文件

通过 FGSM 系列算法生成对抗样本  $\tilde{o}^{k+1} = \tilde{o}^k + \eta_k$

$k = k + 1$

**end**

得到初始样本后, 采用文献[7]、文献[12]或文献[13]的边界攻击算法, 这 3 种边界攻击算法本质上是一致的, 都是基于决策的黑盒攻击手段, 可以通过改变初始样本来减少查询次数, 这里以文献[12]的边界攻击算法为例, 具体算法流程如下所示。

#### 算法 2 新的边界攻击算法

##### Algorithm 2. New boundary attack algorithm

**算法:** 优化初始样本的决策程序算法

**输入:** 目标图像  $o$

**输出:** 对抗样本  $\tilde{o}$ , s.t.  $\min d(o, \tilde{o}) = \|o - \tilde{o}\|_2^2$

**初始值:**  $k = 0$ ,  $\tilde{o}^0$  为上述算法 1 中得到的简易对抗样本, s.t.  $\tilde{o}^0$  和目标图片  $o$  是对抗关系

**While**  $k$  小于设定的最大次数

取一个随机的扰动  $\eta_k \sim P(\tilde{o}^{k-1})$

**if**  $\tilde{o}^k + \eta_k$  和目标图片  $o$  是对抗关系

$\tilde{o}^k = \tilde{o}^{k-1} + \eta_k$

**else**

$\tilde{o}^k = \tilde{o}^{k-1}$

**end**

$k = k + 1$

**end**

### 3.2 无目标攻击流程

根据上述设计的算法, 进行无目标攻击的相关实验, 具体步骤如下:

**步骤一:** 实验样本的准备。选用一个神经网络作为攻击目标 (攻击者并不知道神经网络的内部结构、训练参数和相应算法, 也不知道是采用何种数据集进行训练的, 唯一知道的是此模型是完成图片分类任务的)。这里选择一张目标图片, 将其输入到目标神经网络中, 别被识别为  $o$  类, 然后, 攻击者从百度随机爬取若干  $o$  类的图片和非  $o$  类的图片, 其中下载  $n$  张  $o$  类图片输入目标神经网络模型, 得到  $n-k$  个分类结果为  $o$  类的样本,  $k$  个分类结果为非  $o$  类的样本, 然后随机下载  $n$  张其它类型的图片, 这  $n+k$  个图片作为非  $o$  类的图片样本, 所有图片汇总一起得到  $n-k+1$  个标签为  $o$  的图片样本 (包括 1 张目标图片)、 $n+k$  个标签为非  $o$  类的图片样本, 通过图像旋转、镜像、移位、缩放比例等方法扩充这两类样本, 通过这两类样本训练一个新的神经网络完成二分类任务。

**步骤二:** 获得初始对抗样本。在构建的二分类神经网络中, 采用 M-DI<sup>2</sup>-FGSM 算法计算交叉熵损失的梯度来找出噪声的方向, 生成干扰图像  $\eta$ , 然后将干扰图像加到目标图片样本  $o$  上生成对抗样本, 此时对于新神经网络而言, 对抗样本的生成属于白盒攻击, 然后将这个样本输入到目标神经网络中进行分类识别, 若识别为非  $o$  类, 则就达成了无目标识别任务; 若仍识别为  $o$  类, 则说明目标神经网络的精度很高, 新神经网络二分类后的对抗样本生成也骗不过目标神经网络。此时, 对刚刚生成的对抗样本采取图像旋转、镜像、移位、缩放比例等数据增强技术进行扩充, 然后投入训练集中, 对二分类神经网络增量训练, 生成新的干扰图像  $\eta'$ , 以此反复, 直到目标神经网络得出非  $o$  的结论, 此时, 对抗样本为  $o + \eta_0 + \eta_1 + \dots + \eta_{n'-1}$ , 即叠加了  $n'$  次干扰图像。新的对抗样本即初始对抗样本很大概率将和原图片样本  $o$  差别较大, 不符合对抗样本的要求, 这里将采用新的边界攻击算法进行处理。

**步骤三:** 对抗样本生成。采用文献[12]中的边界攻击算法, 将文献[12]中随机生成的初始样本变更为步骤二得到的简易对抗样本, 以此来进行无目标攻击, 最终得到清晰度符合要求的对抗样本图片。

综上所述, 可以得到整个无目标黑盒攻击流程, 如图 3 所示。



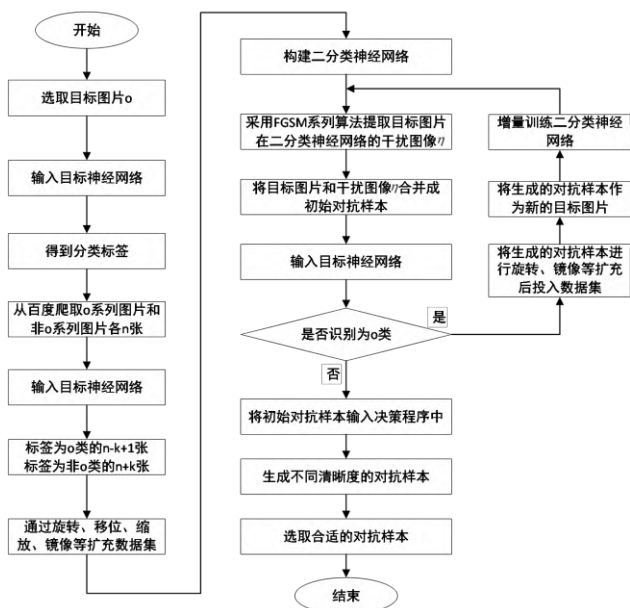


图3 无目标黑盒攻击整体流程

Fig. 3 The whole process of no target black box attack

### 3.3 有目标攻击流程

根据上述设计的算法,进行有目标攻击的相关实验,具体步骤如下:

对于有目标攻击来讲,若按照上述无目标攻击的流程来进行,理论构建的二分类神经网络应变为三分类神经网络,分别为目标图片 $o$ 类、指定错误类 $o'$ 类、还有其他类共3类,再利用例如文献[18-20]等白盒有目标攻击算法来生成初始样本。同样的,然后将这个样本输入到目标神经网络中进行分类识别,若识别为 $o'$ 类,则就达成了有目标识别任务;若仍识别为 $o$ 类或其他类,则说明需要继续生成对抗样本。此时,对刚刚生成的对抗样本采取图像旋转、镜像、移位、缩放比例等数据增强技术进行扩充,然后投入训练集中,对三分类神经网络增量训练,生成新的干扰图像 $\eta'$ ,以此反复,直到目标神经网络得出 $o'$ 类的结论,此时,对抗样本为 $o' + \eta_0 + \eta_1 + \dots + \eta_{n'-1}$ ,即叠加了 $n'$ 次干扰图像。

但实际操作中发现,无目标攻击只需要将对抗样本分为非 $o$ 类即可,而有目标攻击所已方搭建分类模型无论怎么进行增量训练,都和目标神经网络差距太大,准确程度很低,无法准确的将对抗样本粗略的分到 $o'$ 类。因此这里根据文献[7]、文献[12]和文献[13]的边界攻击算法原理,对目标图片 $o$ 和指定错误分类 $o'$ 的一张图片进行加权叠加,目标图片 $o$ 的权重从0.05-0.95,指定分类的一张图片权重从0.95-0.05,步长0.05,如图4所示。



图4 两个类别的加权图

Fig. 4 Weighted graphs of two categories

可以看出,前7幅加权图片被目标神经网络识别为瓢虫(ladybug),其他图片被目标神经网络识别为金毛猎犬(golden retriever),选取合适的加权样本后整体流程和无目标攻击一致,只是将随机扰动的初始样本换为加权样本,通过加权样本来获得合适的初始样本。

理论上第1幅到第7幅图片之间的golden retriever的权值越大,越容易通过算法1来找到合适的初始样本,使得golden retriever错误的识别为ladybug,但是golden retriever的权值越大,初始样本越接近于原始算法直接采用golden retriever作为初始样本的时候,且对查询次数的降低基本没有任何改善;Ladybug的权值越大,就出现前文提到的准确程度很低,无法准确的将对抗样本粗略的分到 $o'$ 类这个问题。

因此,这里选取第1-7幅图片作为加权样本 $m$ 。加权图片越靠近第7幅图片则越接近两个类别的边界,此时仍然通过算法1来进行处理,通过设定每一幅图片的最大尝试查询次数 $p$ ,就容易得到所需要的初始样本。这里的 $p$ 值不需要太大,太大会导致图片严重失真,且不能保证攻击的方向性,这里 $p$ 的取值一般不超过10。最好的情况是选取第一幅图片时就能通过算法1得到合适的初始样本,最差的情况就是到第7幅图片也没有得到合适的初始样本,此时只有使用第8幅图片作为初始样本。

最后再通过算法2来进行无目标攻击,最终得到清晰度符合要求的对抗样本图片。

综上所述,可以得到整个有目标黑盒攻击流程,如图5所示。

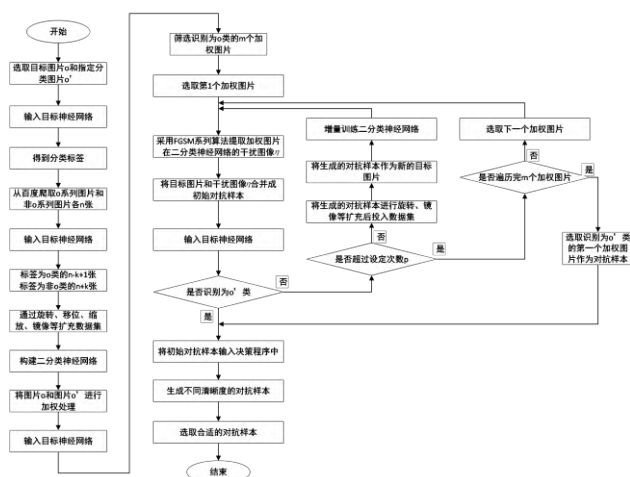


图 5 有目标黑盒攻击整体流程

Fig. 5 The whole process of black box attack with target

## 4 实例分析

### 4.1 无目标攻击实例分析

这里选择目标神经网络为 ImageNet 数据集对应的 ResNet-50 网络结构, 选择目标图片  $o$  为瓢虫 (ladybug) 图片,  $n$  取 100, 根据上节中的步骤, 从百度随机爬取 100 张瓢虫的图片和 100 张其他的图片, 将 100 张瓢虫的图片导入到目标神经网络中, 有 92 张被分类为瓢虫标签, 8 张被分类为其他标签, 此时, 加上目标图片, 共有 93 张瓢虫的图片, 108 张其他的图片, 经过旋转、移位、缩放、镜像等方法, 将数量扩充至 5 倍, 搭建二分类神经网络来训练数据。

采用 M-DI<sup>2</sup>-FGSM 算法计算交叉熵损失的梯度来找出目标图片在新的神经网络噪声的方向, 生成干扰图像  $\eta$ , 合成为对抗样本后输入到目标神经网络中, 发现仍识别为瓢虫, 然后将刚刚合成的对抗样本经过旋转、移位、缩放、镜像等方法进行扩充, 然后作为数据集来增量训练二分类神经网络, 然后再次利用 M-DI<sup>2</sup>-FGSM 算法合成为对抗样本, 经过反复叠加, 7 次叠加后目标神经网络识别为鲤鱼 (crayfish), 如图 6 所示。



图 6 反复叠加后的对抗样本

Fig. 6 Counter sample after repeated superposition

将新的对抗样本输入到边界攻击程序中, 作为起始样本来进行下一步处理, 具体攻击过程如图 7 所示。

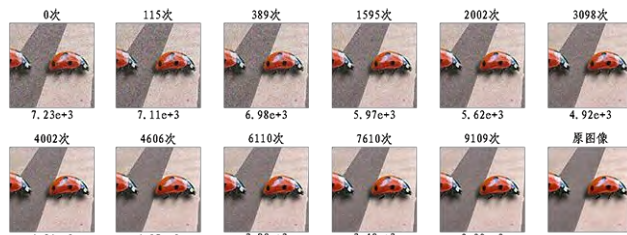


图 7 更改初始样本后的无目标攻击过程

Fig. 7 No target attack process after changing initial sample

可以看出, 当对抗样本与目标图片之间的均方误差约为  $2.99e+3$  时, 共对目标神经网络查询了  $9109+100+7$  次, 此时肉眼已难以区分对抗样本和目标图片之间的差异, 此时目标神经网络仍识别为鲤鱼 (crayfish)。对比第 1 章中的初始样本为随机干扰图片的无目标攻击过程, 当对抗样本与目标图片之间的均方误差约为  $2.99e+3$  时, 共查询了 12006 次。

显然, 更改初始样本后无目标黑盒攻击过程中对目标神经网络的查询次数显著降低, 大约降低了 2790 次, 节省了 23% 的查询次数, 同时, 经过简单的测算, 获得初始样本的过程耗时 (包括爬取图片、查询图片、扩充数据集、训练和增量训练网络、调用 M-DI<sup>2</sup>-FGSM 算法叠加对抗样本) 不超过原来的黑盒攻击查询 2790 次所耗费的时间, 因此, 无目标黑盒攻击算法计算所需时间也在可控范围内。

### 4.2 有目标攻击实例分析

对于有目标攻击, 同样选择目标神经网络为 ImageNet 数据集对应的 ResNet-50 网络结构, 选择目标图片  $o$  为瓢虫 (ladybug) 图片,  $n$  取 100, 根据上节中的步骤, 从百度随机爬取 100 张瓢虫的图片和 100 张其他的图片, 将 100 张瓢虫的图片导入到目标神经网络中, 有 90 张被分类为瓢虫标签, 10 张被分类为其他标签, 此时, 加上目标图片, 共有 91 张瓢虫的图片, 110 张其他图片, 经过旋转、移位、缩放、镜像等方法, 将数量扩充至 5 倍, 搭建二分类神经网络来训练数据。

选择指定分类图片  $o'$  为金毛猎犬 (golden retriever) 图片, 通过 2.3 节中的方法进行加权并按照从第 1 到第 7 的顺序调用 M-DI<sup>2</sup>-FGSM 算法计算交叉熵损失的梯度来多次叠加对抗样本, 设定  $p$  值为 10, 同时将每一步的对抗样本经过旋转、移位、缩放、镜像扩充后作为数据集增量训练二分类神经网络。经过计算, 在第 5 个加权图片中的第 2 次叠加后, 找到了合适的初始样本, 此时对抗样本被识别为金毛猎犬 (golden retriever), 此时已经对目标



神经网络查询了 100+40+2 次。

将新的对抗样本输入到边界攻击程序中,作为起始样本来进行下一步处理,具体攻击过程如图 8 所示。

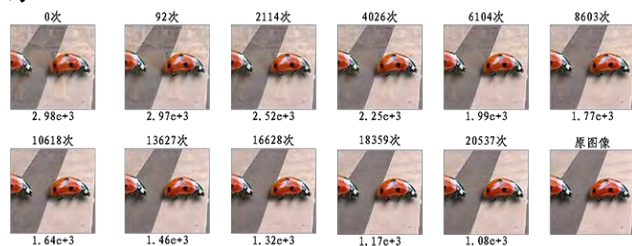


图 8 更改初始样本后的有目标攻击过程

Fig. 8 Target attack process after changing initial sample

可以看出,当对抗样本与目标图片之间的均方误差约为  $1.08\text{e}+3$  时,共对目标神经网络查询了  $20537+100+19+40+2$  次,此时肉眼已难以区分对抗样本和目标图片之间的差异。对比第 1 章中的初始样本为随机干扰图片的有目标攻击过程,当对抗样本与目标图片之间的均方误差约为  $1.08\text{e}+3$  时,共查询了 25058 次。

显然,更改初始样本后有目标黑盒攻击过程中对目标神经网络的查询次数显著降低,大约降低了 4360 次,节省了 17% 的查询次数,同时,经过简单的测算,获得初始样本的过程耗时(包括爬取图片、查询图片、扩充数据集、训练和增量训练网络、加权图像、调用 M-DI<sup>2</sup>-FGSM 算法叠加对抗样本)不超过原来的黑盒攻击查询 4360 次所耗费的时间,因此,有目标黑盒攻击算法计算所需时间也在可控范围内。

## 5 结束语

本文是在基于决策的黑盒攻击算法的基础上,提出可以利用模型的迁移性来循环叠加干扰图像,确定新的初始样本,提高基于决策攻击的起点,降低查询次数。算法有着更好地通用性和普适性,对于各类神经网络都能达到很好地欺骗效果,改进后的算法时间复杂度低,生成对抗样本耗时短,使得对抗攻击更为有效、更稳健、更难以被察觉。通过研究并设计的对抗样本,可以作为神经网络鲁棒性的评估标准,扩展神经网络对抗防御的思路,进一步提高神经网络模型的稳健性;另一方面可以激发更多的神经网络的研究,提高神经网络识别的精度,使神经网络模型更加完善,以达到更好地提升神经网络认知水平的目的。下一步将着重针对边界攻击的过程算法进行优化,采用新的方法对梯度方向进

行估计,对分类边界进行优化搜索,尽可能减少整体的查询时间,提高攻击效率。

## 参考文献 (References)

- [1] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to End Learning for Self-Driving Cars[EB/OL].[2019-07-30].<https://arxiv.org/abs/1604.07316>.
- [2] PARKHI O M, SIMONYAN K, VEDALDI A, et al. A Compact and Discriminative Face Track Descriptor [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE Computer Society, 2014: 1693-1700.
- [3] DONG Y, SU H, WU B, et al. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE Computer Society, 2019: 7706-7714.
- [4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, BRUNA J, ERHAN D, GOODFELLOW I, FERGUS R. Intriguing properties of neural networks[EB/OL]. [2019-07-30].<https://arxiv.org/abs/1312.6199>.
- [5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples [EB/OL].[2019-07-30].<https://arxiv.org/abs/1412.6572>.
- [6] MINHAO CHENG, SIMRANJIT SINGH, PATRICK CHEN, PIN-YU CHEN, SIJIA LIU, CHO-JUI HSIEH. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack[EB/OL]. [2019-07-30].<https://arxiv.org/abs/1909.10773>.
- [7] CHEN J, JORDAN M I, WAINWRIGHT M J. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack[EB/OL]. [2019-07-30].<https://arxiv.org/abs/1904.02144>.
- [8] DONG Y, LIAO F, PANG T, SU H, ZHU J, HU X, LI J. Boosting adversarial attacks with momentum[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE Press, 2018: 9185-9193.
- [9] SARKAR S, BANSAL A, MAHBUB U, CHELLAPPA R. UPSET and ANGRI: breaking high performance image classifiers[EB/OL]. [2019-07-30].<https://arxiv.org/abs/1707.01159>.
- [10] AKHTAR N, MIAN A. Threat of adversarial attacks on



- deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6:14410-14430. <https://arxiv.org/abs/1706.06083>.
- [11] 张嘉楠,王逸翔,刘博,常晓林.深度学习的对抗攻击方法综述[J]. 网络空间安全, 2019, 10(7):87-96.(ZHANG J N, WANG Y X, LIU B, CHANG X L. Survey of adversarial attacks of deep learning[J]. Information Security and Technology, 2019, 10(7):87-96.)
- [12] WIELAND BRENDL, JONAS RAUBER. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models[EB/OL]. [2019-07-30]. <https://arxiv.org/abs/1712.04248>.
- [13] LIU Y, MOOSAVI-DEZFOOLI S M, FROSSARD P. A geometry-inspired decision-based attack[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE Press, 2019: 4889-4897.
- [14] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical Black-Box Attacks against Machine Learning[EB/OL]. [2019-07-30]. <https://arxiv.org/abs/1602.02697>.
- [15] FLORIAN TRAMÈR, NICOLAS PAPERNOT, IAN GOODFELLOW, DAN BONEH, PATRICK MCDANIEL. The space of transferable adversarial examples[EB/OL]. [2019-07-30]. <https://arxiv.org/abs/1704.03453>.
- [16] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. [2019-07-30]. <https://arxiv.org/abs/1607.02533>.
- [17] XIE C, ZHANG Z, ZHOU Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE Press, 2019: 2730-2739.
- [18] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[EB/OL]. [2019-07-30]. <https://arxiv.org/abs/1611.01236>.
- [19] ZHENG T, CHEN C, REN K. Distributionally adversarial attack[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 2253-2260.
- [20] MADRY A, MAKELOV A, SCHMIDT L, TSIPRAS D, VLADU A. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2019-07-30].