计算机科学
COMPUTER SCIENCE

基于差分进化算法的字符对抗验证码生成方法

杨浩，闫巧

引用本文

杨浩，闫巧. 基于差分进化算法的字符对抗验证码生成方法[J]. 计算机科学, 2022, 49(11A): 211100074-5.

YANG Hao, YAN Qiao. Adversarial Character CAPTCHA Generation Method Based on Differential Evolution Algorithm [J]. Computer Science, 2022, 49(11A): 211100074-5.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

开放式环境下基于向量表征与计算的动态访问控制

Vector Representation and Computation Based Dynamic Access Control in Open Environment

计算机科学, 2022, 49(11A): 210900217-7. https://doi.org/10.11896/jsjkx.210900217

基于流量分析发现未知UDP反射放大协议

Discovery of Unknown UDP Reflection Amplification Protocol Based on Traffic Analysis

计算机科学, 2022, 49(11A): 211000089-5. https://doi.org/10.11896/jsjkx.211000089

深度神经网络的对抗攻击及防御方法综述

Survey of Adversarial Attacks and Defense Methods for Deep Neural Networks

计算机科学, 2022, 49(11A): 210900163-11. https://doi.org/10.11896/jsjkx.210900163

对抗性网络流量的生成与应用综述

Generation and Application of Adversarial Network Traffic:A Survey

计算机科学, 2022, 49(11A): 211000039-11. https://doi.org/10.11896/jsjkx.211000039

融合多层次视觉信息的人物交互动作识别

Human-Object Interaction Recognition Integrating Multi-level Visual Features

计算机科学, 2022, 49(11A): 220700012-8. https://doi.org/10.11896/jsjkx.220700012

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100074

基于差分进化算法的字符对抗验证码生成方法

杨 浩 闫 巧
深圳大学计算机与软件学院 广东 深圳 518000
(yanghao20181@email.szu.edu.cn)

摘 要 验证码被广泛应用于网站、应用程序的注册登录环节以区分人类用户与计算机程序。然而随着深度学习的发展,许多针对验证码的深度学习识别方法不断被提出,验证码不再能较好地区分人类用户与计算机程序,验证码的安全性面临着极大挑战。对抗样本可以使神经网络的输出结果产生大幅误差,将对抗样本与验证码结合以抵御深度学习识别系统对验证码的攻击是一种行之有效的方法。将图像领域的对抗样本生成方法用于生成对抗验证码来防御深度学习方法是当前的研究热点之一。现有的字符对抗验证码生成方法都是需要知道攻击网络的结构参数信息的白盒方法,然而在实际的验证码应用场景中通常无法知道攻击网络的信息,健壮性的验证码应该在不知道攻击者信息的情况下依然有良好的防御能力。因此提出了一种基于差分进化算法的黑盒字符型对抗验证码生成方法(Adversarial Character CAPTCHA Generation Method Based on Differential Evolution Algorithm,ACoDE),在无需了解攻击网络信息的情况下通过优化经典差分进化算法变异过程中的缩放因子以及种群进化策略来提高算法的求解能力,使对抗样本误导神经网络的能力更强。将该对抗样本生成方法用于字符验证码数据集后目前最先进的基于卷积神经网络的字符型验证码识别系统的识别准确率降低到了30%以下,且对抗验证码的视觉效果比其他白盒方法生成的对抗验证码更好。

关键词:深度学习;对抗样本;差分进化算法;验证码;网络安全
中图法分类号 TP391

Adversarial Character CAPTCHA Generation Method Based on Differential Evolution Algorithm

YANG Hao and YAN Qiao
School of Computer Science and Software of Engineering,Shenzhen University,Shenzhen,Guangdong 518000,China

Abstract CAPTCHA is widely used in the registration and login process of websites and applications to distinguish normal users from programs. However,with the advancement of deep learning,many deep learning recognition methods for CAPTCHA have been proposed. CAPTCHA can no longer distinguish human users from computer programs effectively,and its security has been greatly challenged. The adversarial example can make the output result of neural network completely different from its original predicted result. Recent researches find that combining adversarial example with CAPTCHA is an effective method to resist the attack of deep learning recognition system. Researchers use adversarial example generation methods to generate adversarial character CAPTCHA to defend against deep learning methods. Existing adversarial character CAPTCHA generation methods are white-box methods that require knowledge of the structural parameter information of the attacking network. However,practical CAPTCHA application scenarios usually do not know the information of the attacking network,so robust CAPTCHA should be able to perform well without knowing the attack information. In this paper,a character-based adversarial CAPTCHA generation method(ACoDE) based on differential evolution algorithm is proposed to improve the solving ability of the algorithm by optimizing the scaling factor in the mutation process and the population evolution strategy. Without knowing the information of the attacking network,the adversarial examples generated by the proposed method are more capable of misleading the neural network. The adversarial example generation method is used for the character CAPTCHA dataset,and the success rate of the current state-of-the-art CNN character-based CAPTCHA recognition system reduce to less than 30%. The visual effect of the adversarial CAPTCHA is satisfactory when compare with other white-box methods.

Keywords Deep learning,Adversarial examples,Differential evolution algorithm,CAPTCHA,Network security

基金项目:国家自然科学基金(61976142);深圳市基础研究面上项目(JCYJ20210324093609025)
This work was supported by the National Natural Science Foundation of China (61976142) and Shenzhen Basic Research Program (JCYJ20210324093609025).
通信作者:闫巧(yanq@szu.edu.cn)

1 引言

验证码(Completely Automated Public Turing Test to Tell Computers and Humans Apart,CAPTCHA)作为一种防御系统被广泛应用于网站、应用程序,是一种重要且普遍的安全保护措施,能有效防止计算机程序进行破解密码、自动刷票、自动发帖等恶意行为。验证码通常会设置一项交互任务,完成这项任务对人类来说较为简单,但是对计算机程序来说却是难题,比如基于文字/数字识别的文本验证码、基于图像识别的图像验证码等。然而深度学习的发展使得这些问题对于计算机来说已经不算难题,如卷积神经网络(Convolutional Neural Networks,CNN)模型识别图像的能力甚至已经超越了人类。已经有很多通过深度学习模型成功破解验证码的研究证明传统的验证码系统已经不安全。字符型文本验证码是最早被发明且应用的验证码之一,用户识别图片后输入字符序列通过验证。它交互方式简单且易于设计,被应用在大多数需要身份认证的领域,然而随着深度学习的发展,字符验证码的安全性难以保障。深度学习的方法基本已经破解了数字与字母组合的字符验证码系统,如 Wang 等[1]提出的 CNN 与 Focal loss 组合的识别网络对一般字符验证码达到了 99%的识别成功率。为了防御基于深度学习的验证码识别攻击,验证码设计者采用字符粘贴、多重字符、字符扭曲、干扰线、添加背景噪声等方式来提升验证码的反识别能力,在一定程度上取得了效果,但是也给正常用户的识别造成了困难,用户对于识别验证码花费的时间大幅提升,同时出错的概率也更高。而且对于这些防御方法,攻击者也提出了新的攻击方法。Gao 等[2]提出了将分割模型与识别模型连接的泛型方法,用来识别不同的添加了反识别干扰的验证码数据集,成功率为 36%89%。字符粘贴与重叠的验证码通常能提升验证码的防御能力,但是对于可用性的破坏非常大,然而即便是采取了此策略的 Google CAPTCHA 和 reCAPTCHA[3]也已经被证明不再安全。由于字符验证码的识别与图像的定位分割识别任务相似,研究者考虑将在图像领域表现优秀的对抗样本与验证码结合。FGSM[4]对抗样本生成方法是经典的基于神经网络梯度的对抗样本生成方法,Kwon 等[5]延续该方法的思路提出了迭代的修改梯度的方法用于生成对抗验证码。该方法需要对图像进行全图修改,对抗样本呈现模糊的特点,对于用户的视觉效果仍然不够友好。单像素攻击[6]在图像识别任务中极端情况下仅需要改变一个像素点就能实现对抗攻击,且该方法与识别网络无关,可以在不同网络中实现迁移攻击。我们考虑利用该方法的特性,对单像素攻击方法进行改进后生成对深度学习攻击方法具有较好的防御作用,同时对用户友好的对抗性字符验证码。本文方法的优势在于:1)寻找全局最优解的可能性更高;不同于局部搜索的贪婪策略,全局搜索的求解能力更强;2)无需获取攻击网络的梯度、网络结构、权重等信息,可以实现对未知攻击网络的防御。

本文提出了一种基于差分进化算法的对抗验证码生成方法(ACoDE),该方法能有效防止卷积神经网络对字符验证码的识别攻击。本文的主要贡献有以下两点:1)对单像素攻击使用的经典差分进化算法进行了改进,使用动态的缩放因子和组合的进化策略使算法求解能力更强,生成对抗样本的成功率更高;2)将该方法用于字符型对抗验证码的生成,使其能误导卷积神经网络的识别,抵御基于深度学习的攻击方法,由于对原验证码的修改很小从而使验证码具有较好的可用性,该方法在对抗性字符验证码领域首次应用了黑盒攻击来生成字符对抗验证码。

本文第 2 节介绍本文涉及的相关概念;第 3 节介绍基于差分进化算法的对抗验证码生成方法;第 4 节对实验结果进行分析;最后总结全文并展望未来。

2 相关工作

2.1 基于深度学习方法的文本验证码识别技术

近年来,深度学习方法飞速发展,其对于图像的定位分割识别任务效果出色。文本验证码通常需要先对验证码图片进行分割,然后分类识别。研究者们将深度学习方法应用于文本验证码识别,使得即便是采用了中空、粘贴、扭曲等处理的文本验证码也不再安全。Tang 等[7]提出一种基于条件生成网络(Conditional Generative Adversarial Networks,CGAN)的识别方法,使用两次 CGAN 来生成验证码。在预处理阶段根据真实验证码生成无干扰信息的验证码,第二次对无干扰验证码的字符间距进行拉伸,为分割阶段提供更容易分割的验证码,之后通过 GooLeNet 进行识别。

Cao 等[8]提出一种对抗学习方法来识别文本型验证码。先训练一个 Pix2pix 网络对验证码图片进行预处理,然后对抗训练出一对分割和识别网络。分割网络不仅能分割粘贴字符,而且可以筛选出难以分割的验证码结果。识别网络采用上下文相关的多通道卷积网络,能有效解决分割过程中因信息丢失而无法识别的问题。

Shu 等[9]使用端到端深度 CNN-RNN 网络来识别验证码。该 CNN-RNN 系统首先利用 CNN 网络结构提取输入图像的特征,然后通过 RNN 网络结构进一步提取和对特征进行分类。CNN-RNN 模型对 4 字符验证码的准确率达到 99%。

2.2 对抗样本的原理与生成方法

Szegedy 等[10]在图像领域首次发现了对抗样本,其仅仅对图像进行很小的扰动就能使分类器无法正确识别图像。通常来说这种扰动十分微小以至于人类无法察觉。他们将一张熊猫的照片的图片样本输入分类器,分类器以 57.7%的置信度将其判断为熊猫。然而,在添加一个难以察觉的扰动后,分类器判断它为一只长臂猿的置信度为 99.3%。结果表明,这种对数据集添加微小的干扰的攻击方式会完全改变神经网络分类器对图像的预测。目前对抗样本的方法较多,最常见的几种方法包括快速梯度符号法、单像素攻击法和通用对抗扰动。

(1)快速梯度符号法

Goodfellow[4]提出了快速梯度符号法(Fast Gradient Sign Method,FGSM)来生成对抗样本,通过不断修改梯度方向来控制扰动的 L-∞ 距离以生成对抗样本。通常分类模型在更新参数时都要使损失值越来越小,从而使模型预测的成功率增加,而 FGSM 方法在更新时,计算梯度方向后在输入图像上增加梯度,使损失越来越大,从而达到攻击的效果。

(2)单像素攻击

Su 等[6]提出了单像素攻击方法(One-pixel Attack),这是

一种极端的对抗攻击方法，仅改变图像中的一个像素值就可以实现对抗攻击。Su 等使用了差分进化算法，对每个像素进行迭代地修改生成子图像，并与母图像对比，根据选择标准保留攻击效果最好的子图像，实现对抗攻击。这种对抗攻击不需要知道网络参数或梯度的任何信息。

(3)通用对抗扰动

通用对抗扰动[11]是指将同一个扰动加入到不同的图片中，能够使图片被分类模型误分类，无论该图片的原类型是什么。作者提出一种算法来寻找一个对抗扰动，该扰动可以批量添加在所有实验样本上，批量地产生对抗图像。实验证明这种扰动还可以泛化到其他网络上。

2.3　对抗样本在验证码中的应用

Osadchy 等[12]在图像验证码领域提出了 DeepCAPTCHA，它采用精心设计的对抗样本来欺骗深度学习图像分类系统。大多数情况下，图像的对抗性攻击对降噪、过滤或其他去除对抗噪声的方法较脆弱，但 DeepCAPTCHA 可以产生不可变的对抗样本。在他们的讨论中，由于 FGSM 对过滤防御方法的鲁棒性较差，因此提出了一种迭代版本的 FGSM。给原图像一个目标标签和一个置信度，并保证产生对抗样本被分类为目标标签和所需的置信度(同时保持添加的噪声最小)；然后在向目标标签靠近的方向迭代运行噪声生成的步骤，直到达到目标标签和所需的置信度级别。

Shekhar 等[13]首先提出了可以破坏没有背景噪音或中等背景噪音的任何数量的语音数字的攻击系统，该系统能够达到 99%～100%的准确性。之后他们使用对抗样本生成算法来生成对抗性音频数据。对模型进行再训练后攻击精度降低，攻击者的成功率只有 25%～36%。作者对比了不同的生成对抗样本的算法，如基本迭代法(BIM)和 DeepFool 音频验证码。结果发现只要攻击者从各种对抗音频数据集中获得的样本少于 45%，对抗样本防御方法就可以成功地防止攻击。

Hitaj 等[14]提出了 CAPTURE 方法，通过在图片上添加对抗样本块的方法来增强验证码图片的鲁棒性。他们将深度神经网络无法识别的无意义图块作为对抗样本添加到图片验证码上，使图片验证码能够抵御深度神经网络的识别。然而该方法的效果与对抗样本块占据图片的比例有关，当对抗样本块的比例超过 50%时才能达到 95%以上的防御效果，而这种验证码图片极大地影响用户的使用，验证码的可用性较差。

Kwon 等[5]提出了对抗性字符验证码的生成方案，采用了 FGSM，I-FGSM，DeepFool 这几种图片对抗样本生成方法。在特定的训练参数和迭代次数后，识别模型对 FGSM 和 I-FGSM方法生成的对抗验证码图片的识别率降低为0%，对 DeepFool 方法生成的验证码图片的识别率降低为 45%。但是该识别模型在测试集上的识别准确率只有 71.42%，通常验证码破解系统能够达到 90%以上的识别准确率，该方案对于当前效果最好的验证码识别系统的防御能力还未被证明。

2.4　差分进化算法

差分进化算法(Differential Evolution，DE)[15]是由 Das 等于 1997 年提出的一种基于群体差异的启发式并行搜索方法，提出的初衷是为了求解切比雪夫多项式问题。差分进化算法被应用于求解高维空间下的多目标优化问题，经典 DE 算法包括种群初始化、变异、交叉以及选择等操作。

差分进化算法基于进化策略，待求解问题被称为目标函数，用 $f(C)=(c_1,c_2,\cdots,c_D)$表示，满足 $L<C<U$，$(c_1,c_2,\cdots,c_D)$是一组待确定的参数，其中 D 代表解空间维度。DE 算法首先初始化覆盖上下界为(L,U)的解空间的种群 $NP=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N)$，种群中包含 N 个候选解。从种群中随机选取 $\boldsymbol{X}_{r1},\boldsymbol{X}_{r2},\boldsymbol{X}_{r3}$进行如下差分变异运算：

$$\boldsymbol{U}=\boldsymbol{X}_{r1}+F(\boldsymbol{X}_{r2}-\boldsymbol{X}_{r3}) \tag{1}$$

其中，F 为缩放因子，它通过控制差分变异操作来影响算法的探索能力与收敛速度。差分变异产生变异向量 $\boldsymbol{U}$，$\boldsymbol{U}$ 与种群中的最优向量竞争，选择表现更优的向量进入下一代种群，重复此过程直到满足目标函数的约束条件或达到最大迭代次数。

随着研究的发展，差分策略有了更多形式，通常 DE 算法的差分策略可以表示为 DE/x/y/z，其中参数 $\boldsymbol{x}$ 表示参与变异的基向量，可以是随机向量(rand)、当前种群的最优向量(best)或者是当前向量本身(current)；参数 y 表示参与变异的差分向量数目；参数 z 表示交叉的模式，如二项式交叉、指数交叉。式(1)描述的经典差分算子表示为 DE/rand/1，基向量为随机选取，差分向量数量为 1。一些常用 DE 算子[15]如式(2)－式(4)所示：

$$\text{DE/rand/2}:\boldsymbol{U}=\boldsymbol{X}_{r1}+F(\boldsymbol{X}_{r2}-\boldsymbol{X}_{r3})+F(\boldsymbol{X}_{r4}-\boldsymbol{X}_{r5}) \tag{2}$$

$$\text{DE/best/1}:\boldsymbol{U}=\boldsymbol{X}_{\text{best}}+F(\boldsymbol{X}_{r1}-\boldsymbol{X}_{r2}) \tag{3}$$

$$\text{DE/current-to-best/1}:\boldsymbol{U}=\boldsymbol{X}+F(\boldsymbol{X}_{\text{best}}-\boldsymbol{X})+F(\boldsymbol{X}_{r1}-\boldsymbol{X}_{r2}) \tag{4}$$

3　基于差分进化算法的对抗样本生成方法

验证码需要具备两个特性，安全性与可用性。安全性是指验证码能够抵御深度学习方法在内的攻击手段的破解，使计算机程序无法轻易地破解验证码系统；可用性是指验证码交互应该简易方便，用户可以在较短时间内识别出验证码并完成验证码输入。一个优秀的验证码系统应该同时具备可用性与安全性。然而由于深度学习的威胁性太强，许多验证码不得不设计得复杂难辨来抵御验证码识别系统，以保证安全性，如图 1 所示。然而在使用了诸如字符扭曲、粘粘干扰线等方法处理后，虽然对于深度学习识别方法的防御能力更强，但验证码图片却变得难以识别，用户需要花费更多的时间来识别字符，甚至还会错误识别验证码。这无疑是破坏了验证码的可用性。

图 1　使用了干扰线的字符验证码

Fig. 1　Character CAPTCHA picture with interfering lines

使用对抗攻击来生成对抗验证码可以降低对验证码图片的修改程度，但是当前对抗字符验证码的生成方法主要是白盒方法，需要根据识别网络的参数来计算对图片添加的扰动，而且处理过的验证码图片清晰度会大幅降低。在实际场景中，识别模型种类繁多，且一般无法了解验证码识别系统的信息，故黑盒方法更具有实际意义。

差分进化算法在全局优化问题中有优秀的表现，它通过模拟生物进化过程反复迭代保留最适应环境的最优个体达到

寻找最优解的目的。Su 利用此思想在图像识别领域提出的单像素攻击仅仅需要改变一个或几个像素点就能实现对深度识别模型的对抗攻击，从视觉上来说几乎无法发现对原图像的改动。

Su 将图片的对抗样本生成问题形式化地描述为一种带有约束的优化问题。将输入的图片用一个向量表示，该向量中的每一个标量元素代表一个像素。图片分类器 f 会将输入的图片 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 正确分类为类别 t，$\boldsymbol{X}$ 属于类别 t 的置信度为 $f_t(\boldsymbol{X})$。$e(x)=(e_1,\cdots,e_n)$ 是添加在 $\boldsymbol{X}$ 上的对抗扰动，最大边界为 L，L 受 $e(x)$ 向量维度的约束。在目标攻击条件下该问题的目的是搜寻满足以下约束的最优解 $e(x)^*$，在非目标攻击条件下该问题描述为[6]：

$$\begin{aligned}&\text{maximize } e(x)^* \ f_{adv}(x+e(x)^*)\\&\text{s.t. } \|e(x)*\|_0\leqslant d\end{aligned}\tag{5}$$

其中，d 代表修改的像素点的数量，他是一个较小值，在原实验中分别为 1，3，5。

该方法相较于一些经典的对抗样本生成方法，如 FGSM，成功率还不够高。在 FGSM 方法达到 90%以上的成功率的情况下，单像素攻击只能达到 70%左右的攻击成功率。

3.1 优化差分进化算法的黑盒对抗验证码生成方案

针对上述问题，我们提出了通过优化差分进化算法来生成字符对抗验证码的黑盒方案 ACoDE。差分进化算法求解能力与待求解问题的特征信息无关，与自身的变异策略、差分策略等相关。ACoDE 仅依赖于验证码识别模型的输出，通过优化变异参数与变异策略来寻找使模型输出错误结果的像素点的位置，在对原图片仅进行少量修改的条件下完成验证码的生成，解决了在黑盒环境下生成对抗验证码的问题，且保证了用户可用性。

差分进化算法的变异阶段对算法性能的影响最大，其通过缩放因子 F 和变异策略来调整算法性能。一般来讲，在进化过程的早期，DE 算子需要一个较大的缩放因子 F 来维持种群的多样性。随着进化的继续，在进化过程的后期，种群中个体向全局最优解收敛，应该采用较小的 F 来加快收敛速度。基于这种思想，我们设计了动态的缩放因子 DF，随着进化次数逐渐缩小，F 的变化范围为(0.95，0.5)，其描述如下：

$$DF=1/(1+e^{(-2.2*(1-g/G))})\tag{6}$$

其中，g 为当前进化代数，G 为最大进化代数。该函数随进化代数的增加呈平稳下降趋势，如图 2 所示。

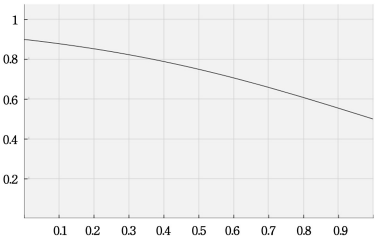


图 2 动态缩放因子随着进化代数的变化趋势

Fig. 2 Trend of dynamic scaling factor with evolutionary generations

接下来我们将叙述 ACoDE 的基本流程。

3.1.1 初始化

将扰动编码为一个数组并作为 DE 算法的候选解。一个候选解包含固定数量的扰动，每个扰动是一个五元组：x，y 坐标以及 RGB 值。利用均匀分布对初始种群进行初始化，以此生成 x-y 坐标。RGB 值由高斯分布 $N(\mu=128,\sigma=127)$ 初始化。

3.1.2 变异

种群在第 g 代进化过程中利用变异算子对个体 $\boldsymbol{X}$ 进行变异，产生下一代变异向量 $\boldsymbol{X}(g+1)$。当变异策略为经典差分进化策略时，每一代种群中的 N 个候选解通过下式产生子代：

$$x_i(g+1)=x_{r1}(g)+DF(x_{r2}(g)-x_{r3}(g))\tag{7}$$

其中，$r1$，$r2$，$r3$ 各不相等，x_i 是候选解中的元素。一个扰动对应修改的一个像素。一旦生成，每个候选解与其对应的父代竞争优胜者优先策略生成下一个迭代。

在求解问题的不同阶段，不同的变异算子所表现出的能力也各不相同。根据算法不同阶段采取不同的变异策略可以有效提升算法性能。我们根据进化代数将算法求解过程分为两个阶段，当进化代数小于最大进化代数一半时为进化过程的第一阶段，需要算法具备更好的探索能力，采用 DF 变异系数的 DE/rand/2/算子：

$$\begin{aligned}x_i(g+1)=&x_{r1}(g)+DF(x_{r2}(g)-x_{r3}(g))+\\&DF(x_{r4}(g)-x_{r5}(g))\end{aligned}\tag{8}$$

进化代数大于最大进化代数一半时为第二阶段，需要更好的局部搜索能力，采用 DF 变异系数的 DE/current-to-best/1/算子：

$$\begin{aligned}x_i(g+1)=&x_{r1}(g)+DF(x_{best}(g)-x_{r1}(g))+\\&DF(x_{r2}(g)-x_{r3}(g))\end{aligned}\tag{9}$$

3.1.3 计算适应度函数

适应度函数即为识别模型对输入个体的识别置信度，对于输入的图片 $\boldsymbol{X}$，分类器 f 会以一定的置信度将其分类，而算法需要找到修改的扰动图片输出使该类别的置信度最低，若满足条件则输出对抗样本，若不满足则进行下一次进化过程。

3.2 生成字符对抗验证码

完成算法的优化设计后我们用该方法来生成对抗验证码。首先分析用于生成对抗样本的字符验证码图片。验证码图片的大小为 128 * 64 像素，图片中包含从集合 $C=\{0,1,\cdots,9,A,B,\cdots,Z\}$ 共 36 个字符中随机选取的 4 个字符。验证码采取添加随机数量干扰线的方式增加其对一般神经网络的抵抗能力。接下来采用 3.1 中的方法生成对抗验证码，如算法 1 所示。

算法 1 基于差分进化算法的对抗验证码生成方法

输入：($\mathbf{X}$，G，N，w，h，f)

输出：($\mathbf{X}_{adv}$)

1. Initial P with(G，N，w，h) //根据最大进化次数 G，种群规模 N，图片参数 w，h 初始化种群 P
2. While g<G Do
3. Select Xi from P (X_1，…X_N)
4. If |f (x_i+e(x_i) *)−f(x)|>α //选择种群中的最优解，计算适应度函数
5. Outputxadv=x+e(x) * //如果满足适应度函数，输出 e(x) *
6. Else
7. For i:N←1
8. F=1 / (1+$e^{(-2.2*(1-g/G))}$)
9. IF statement 1
10. $x_i(g+1)=x_{r1}(g)+DF(x_{r2}(g)-x_{r3}(g))+DF(x_{r4}(g)-x_{r5}(g))$

11.　IF statement 2 //根据进化阶段选择进化策略

12.　$x_i(g+1)=x_{r1}(g)+DF(x_{best}(g)-x_{r1}(g))+DF(x_{r2}(g)-x_{r3}(g))$

13.　g ← g+1，update P

14.　END For

15. END While

16. END

该方法可以为一张图片生成对抗验证码，图 3 描述了 ACoDE 生成对抗验证码并攻击 CNN 识别模型的过程。我们将数据集图片依次应用本方法，结果表明生成的对抗性验证码视觉效果良好易于识别，用户可以清楚快速地完成验证交互。

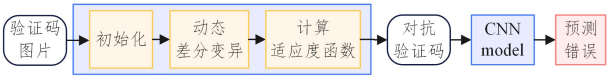


图 3　ACoDE 的整体流程

Fig. 3　Overall process of ACoDE

4　实验结果与分析

4.1　在 CIFAR-10 数据集上的结果

Su 的方法中基本差分进化算法缩放因子 $F=0.5$，变异策略为 DE/best/1/bin。实验攻击的识别网络为全卷积网络 PureCNN 与 network in network(NIN)网络，在 CIFAR-10 数据集上训练的准确率为 88.8%与 90.8%，分别改变 1，3，5 个像素点进行攻击，本文复现了作者的方法，并统计了 500 张图片的攻击结果。

对于 network in network 网络，攻击方法改变的像素点分别为 1，3，5 时成功率分别为 34%，65.8%，65.4%；对于 PureCNN 网络，在改变的像素点为 1，3，5 时成功率分别为 17%，44%，48%，随着改变的像素点增加，攻击能力更强。以上结果基本符合原论文中的描述，该组实验数据将作为之后我们改进方法的对照组数据。我们将采用动态的缩放因子，变异策略不变的 DE 算法命名为 ACoDE1，将采用动态缩放因子与组合变异策略的 DE 方法命名为 ACoDE2。采用我们提出的改进方法后的效果如表 1 所列。

表 1　改进的差分进化算法不同情况下的成功率

Table 1　Success rate of improved differential evolution algorithm in different cases

network/Number of fixed pixel	One-pixel attack[6]	ACoDE1	ACoDE2
NIN/1	0.34	0.35	0.38
NIN/3	0.69	0.61	0.69
NIN/5	0.66	0.67	0.72
PureCNN/1	0.17	0.19	0.19
PureCNN/3	0.44	0.46	0.49
PureCNN/5	0.48	0.44	0.51

实验结果表明，大多数情况下 ACoDE1 相较于原方法在两个网络上的效果更优，ACoDE2 在所有情况下都要优于原方法与 ACoDE1，证明了 ACoDE2 生成对抗样本的成功率更高。实验结果中出现了 ACoDE1 的部分效果比 one-pixel attack 差的现象，这是因为差分进化算法的效果与差分策略、相关系数选取、进化次数等都有关系，理论上进化次数越多搜寻全局最优解的能力越强。但实际中由于计算条件的限制，最大进化次数设置为 1 000 次，有可能出现达到最大进化次数时种群还未达到加速收敛的情况。ACoDE1 相比于原方法在进化初期加强了探索能力，后期加强了收敛能力而减弱了探索能力，在 1 像素任务中具有更好的效果。而在 3，5 像素任务中由于后期减弱了探索能力导致达到最大进化次数时部分效果下降。在 ACoDE1 基础上采用组合变异策略的 ACoDE2 兼顾了探索能力与收敛能力，在对抗样本生成问题中表现更优。

4.2　生成字符型对抗验证码的效果

接下来我们用 ACoDE2 方法在字符验证码图片数据集上生成对抗验证码。实验数据集为 python 生成。我们生成了 20 000 张 4 字符验证码，验证码的文本内容在阿拉伯数字 0－9 与英文大写字母 A－Z 共 36 个字符中随机选取，将数据集的 90%划分为训练集，10%为测试集。用于识别字符验证码的模型是两个卷积神经网络，CNN1 是 3 个卷积-池化模块的简单网络，CNN2 是深度学习破解验证码竞赛中表现优异的深度网络。在 Keras 框架下构建模型，经过训练后网络对于字符验证码的识别准确率分别为 90.2%与 98.6%。

我们分别设置像素点的数量 d 为 1，3，5 来检验算法生成对抗验证码的性能。当识别网络对 4 字符验证码中任意一个字符识别错误即为攻击成功。对 CNN1 网络攻击的成功率分别为 33%，66.5%，78.6%，对 CNN2 网络攻击的成功率分别 24.1%，63.2%，72.8%，如图 4、图 5 所示。

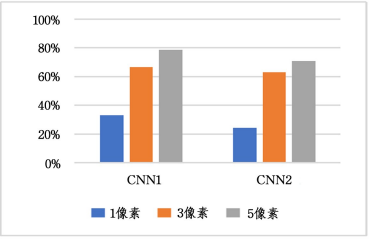


图 4　d 设置为 1，3，5 时 ACoDE 方法的攻击效果

Fig. 4　Attack effect of ACoDE method when $d=1,3,5$

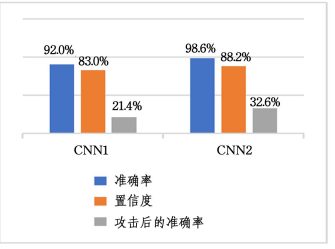


图 5　攻击前后识别模型的准确率对比

Fig. 5　Accuracy comparison of recognition models

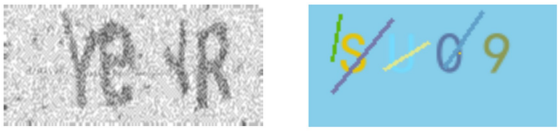
相比其他对抗样本生成方法，本文方法对原验证码的修改比例非常小，仅仅在个位数的像素点数，完全不会影响用户对验证码的识别，而 FGSM 方法与 UAP 方法都需要对图片进行整体修改，CAPTURE 方法需要修改超过 50%的原图片比例才能达到最优效果，在视觉效果上显得模糊，需要用户花费更多时间来辨认验证码。表 2 对比了几种对抗样本生成方法的修改比例。图 6 对比了 FGSM 与本文方法生成的验证码的视觉效果，左边 FGSM 方法生成的验证码改变了原图片的背景色与字符颜色，视觉效果模糊，右边为 ACoDE 生成的验证码，仅改变了一个像素。可以直观看出本文方法生成的

对抗验证码的可用性更好。

表 2 ACoDE 与其他方法的对比

Table 2 ACoDE versus other methods

Generation method	Percentage of adversarial example/%	Black/White-box method
ACoDE(proposed method)	0.48	Black-box
FGSM	100	White-box
UAP	100	White-box
CAPTURE	20～60	White-box



(a)FGSM 方法 (b)ACoDE 方法

图 6 FGSM 方法[5]与 ACoDE 生成的对抗验证码的视觉效果

Fig. 6 Visual effect of adversrial captcha generated by FGSM and ACoDE

结束语 本文基于单像素对抗样本生成方法提出了一种改进的字符对抗验证码生成方法 ACoDE。改进了单像素攻击方法的缩放因子与变异策略，在 CIFAR-10 图片数据集上的实验结果表明改进后的方法比原方法生成对抗样本的成功率更高。我们还将此方法用于字符型验证码数据集，生成了能够抵御卷积神经网络识别的对抗性字符验证码，达到 78.6%的成功率。同时我们生成的对抗验证码相较于其他验证码生成方法生成的验证码对原样本的改动非常小，在视觉效果上更加优秀，验证码对用户的可用性更好，且本文方法是首次应用黑盒攻击来生成对抗验证码的方法。之后的研究将在两个方面进行扩展，一是对提高 DE 算法的求解能力进行研究，在 DE 算法领域已经有很多工作，对于初始化方法，研究者们对缩放因子的优化以及差分策略设计等都有深入的研究，如果能够将这些优化方法，应用于我们提出的 ACoDE 验证码生成方法还会有更多提升空间；二是将本文方法应用在其他类型验证码，如图像验证码、语音验证码、视频验证码等领域，沿用本文中的研究思路，将对抗样本的生成问题转化为在原样本中寻找使得识别模型输出错误结果的微小样本问题，并用差分进化算法对该问题求解，使所提方法更具有普遍性。

参 考 文 献

[1] WANG Z, SHI P. CAPTCHA Recognition Method Based on CNN with Focal Loss[J]. Complexity, 2021(2): 1-10.

[2] GAO H C, WANG W, FAN Y, et al. The Robustness of "Connecting Characters Together" CAPTCHAs[J]. Journal of Information Science and Engineering, 2014, 30(2): 347-369.

[3] VON AHN L, MAURER B, MCMILLEN C, et al. Recaptcha: Human-based character recognition via websecurity measures [J]. Science, 2008, 321(5895): 1465-1468.

[4] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

[5] KWON H, YOON H, PARK K W. Robust CAPTCHA Image Generation Enhanced with Adversarial Example Methods[J]. IEICE Transactions on Information and Systems, 2020, E103-D(4): 879-882.

[6] SU J, VARGAS D V, KOUICHI S. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[7] TANG Z Y, TIAN C X, LI J, et al. A text-based CAPTCHA recognition method based on conditional generative adversarial networks[J], Chinese Journal of Computers, 2020, 43(8): 199-204.

[8] CAO Y R, LU L, GONG Y H, et al. A Captcha Recognition Method based on Adversarial Network[J]. Computer Engineering and Applications, 2020, 56(8): 199-204.

[9] SHU Y, XU Y. End-to-End Captcha Recognition Using Deep CNN-RNN Network[C]// 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference(IMCEC). 2019: 54-58.

[10] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]// International Conference on Learning Representations. 2014.

[11] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 86-94.

[12] OSADCHY M, HERNAN DE Z-CASTRO J, GIBSON S, et al. No Bot Expects the DeepCAPTCHA! Introducing Immutable AdversarialExamples, With Applications to CAPTCHA Generation[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2640-2653.

[13] SHEKHAR H, MOH M, MOH T. Exploring Adversaries to Defend Audio CAPTCHA[C]// 2019 18th IEEE International Conference On Machine Learning And Applications(ICMLA). 2019: 1155-1161.

[14] HITAJ D, HITAJ B, JAJODIA S, et al. Capture the Bot: Using Adversarial Examples to Improve CAPTCHA Robustness to Bot Attacks[J]. Intelligent Systems, IEEE, 2020, 36(5): 104-112.

[15] DAS S, SUGANTHAN P N. Differential Evolution: A Survey of the State-of-the-Art[J]. IEEE Transactions on Evolutionary Computation, 2011, 15(1): 4-31.

YANG Hao, born in 1995, postgraduate. His main research interests include network security and machine learning.

YAN Qiao, born in 1972, Ph.D, professor, is a member of China Computer Federation. Her main research interests include network security, software-defined networking and adversarial machine learning.