一种基于条件生成式对抗网络的文本类验证码识别方法

汤战勇[1),2)] 田超雄[1),2)] 叶贵鑫[1),2)] 李 婧[3)] 王 薇[1),2)]
龚晓庆[1),2)] 陈晓江[1),2)] 房鼎益[1),2)]

[1)](西北大学信息科学与技术学院 西安 710127)
[2)](陕西省无源物联网国际联合研究中心 西安 710127)
[3)](陕西理工大学数学与计算机科学学院 陕西 汉中 723001)

摘 要 验证码被广泛应用于网站登录、注册等环节,用来增强身份验证和防止来自计算机程序的自动攻击.其中文本类验证码由于密码空间大、交互方式简单等特点被大多数主流网站使用.目前,为了增加计算机程序对文本类验证码自动识别的难度,设计时普遍将复杂干扰信息、字符扭曲、旋转和粘连、不同类型字体等安全性特征随机组合使用.由于组合了多种安全特征,传统的验证码识别方法对该种验证码的识别率非常低甚至失效.针对此类文本类验证码,本文提出了一种基于条件生成式对抗网络(CGAN)的通用识别方法.该方法利用 CGAN 去除验证码中的背景干扰信息并拉伸验证码中的字符间距,以生成无干扰且无字符粘连的验证码.然后使用本文优化组合的分割算法对验证码进行有效分割,再通过 GoogleNet 对分割后的单个字符进行识别.并且在难以以低成本大量获取真实验证码的情况下,本文设计了程序模拟验证码对网络进行训练,训练成本远低于现有其他方法且训练效果良好.最终的实验结果表明,本文提出的方法能够成功的识别 Microsoft、Wikipedia、百度、支付宝、新浪等国际主流网站的验证码,识别率相较于传统方法最大提升度可达到 70.2%.

关键词 文本类验证码;验证码识别;条件生成式对抗网络;字符分割;去干扰算法
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.01572

A Recognition Method for Text-Based Captcha Based on CGAN

TANG Zhan-Yong[1),2)] TIAN Chao-Xiong[1),2)] YE Gui-Xin[1),2)] LI Jing[3)] WANG Wei[1),2)]
GONG Xiao-Qing[1),2)] CHEN Xiao-Jiang[1),2)] FAND Ding-Yi[1),2)]

[1)](*School of Information Science and Technology, Northwest University, Xi'an* 710127)
[2)](*Shaanxi International Joint Research Centre for the Battery-free Internet of Things, Xi'an* 710127)
[3)](*School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, Shaanxi* 723001)

Abstract Captchas are widely used in the login and registration of websites to enhance authentication and prevent automatically attacks from computer programs. Captchas can be divided into three categories: text-based captcha, image-based captcha and audio-based captcha. Among the three captcha schemes, text-based Captchas are extensively used by most mainstream websites for its large password space and simple interaction mode. Due to the wide deployment of

收稿日期:2019-02-21;在线发布日期:2019-09-23.本课题得到国家自然科学基金(61672427,61972314)、陕西省国际合作计划(2017KW-008)、陕西省国际合作计划(2019KW-009)、陕西省重点研发计划(2017 GY-191)、陕西省创新团队(2018SD0011)资助.**汤战勇**,博士,教授,中国计算机学会(CCF)会员,主要研究方向为软件安全与保护、无线传感网安全.E-mail:zytang@nwu.edu.cn.**田超雄**,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为信息安全.**叶贵鑫**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为软件自动检测与漏洞挖掘、身份认证.**李 婧**,硕士,讲师,中国计算机学会(CCF)会员,主要研究方向为软件安全.**王 薇**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为移动计算、网络与信息安全.E-mail:wwang@nwu.edu.cn.**龚晓庆**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为软件安全、工作流.**陈晓江**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为无线传感网络、软件安全与保护.**房鼎益**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为网络与信息安全、软件安全与保护、无线传感器网络关键技术.

text-based captchas, a compromise on the captcha scheme can have significant implications and could result in serious consequences. At present, in order to protect text-based captcha against automatic recognition by computer programs, text-based Captchas generally use a random combination of different security features, such as complexity obstacle backgrounds, characters warp, rotate and overlapping. In spite of researchers have proposed a number of attacks, text-based Captchas are still being used by most popular websites such as Google, Microsoft, Alipay, etc. One of the reasons is that the previous model-based attacking methods are scheme-specific. This means that a small change in the captcha such as a noisier background, more overlapping characters can easily invalid a prior attack. The other reason is that prior deep learning-based attacks require millions of training samples. However, collecting and labelling so many captchas requires a labor-intensive and time-consuming process to construct. In order to address the above challenges, we present a generic generative adversarial network-based attack on text-based Captchas. Unlike previous machine-learning-based attacks, our approach does not require a large volume of real captchas to learn an effective solver, we significantly reduces the number of real captchas needed. This is achieved by first using CGAN to remove the background interference information and stretch the character spacing of the Captchas. Then we use the optimization of segmentation algorithms to segment the stretched Captchas effectively, and use GoogLeNet to perform single character identification. In addition, it is difficult to obtain a large number of real Captchas at a low cost. This paper designs a program to simulate these real Captchas for network training, and the training cost is far lower than other existing methods and the training effects are the same. In our experiments, we automatically collected 10 text-based captcha schemes which are widely used by some of the top-50 popular websites ranked by alexa.com as of March, 2018. These websites include eBay, Alipay, JD, Wikipedia, many of which employ advanced security features such as complex background, rotated, distorted and overlapping characters. For each captcha scheme, we collected and labelled 200 target captchas. We generated 30000 synthetic captchas for training the preprocessing model and the captcha solver. We evaluate our approach using the collected captchas. The experimental results demonstrate that our generic attack needs less number of real text-based captchas instead of millions to learn a text-based captcha solver, but the learned captcha solver can greatly outperform prior start-of-the arts. Extensive experimental results also show that the method proposed in this paper can successfully identify the Captchas catch from some famous websites, such as Microsoft, Wikipedia, Baidu, Alipay, Sina, and so on. Furthermore, the highest success rate of our approach is 70.2% higher than the traditional methods.

Keywords　text-based CAPTCHAs; CAPTCHAs identification; conditional generative adversarial networks; character segmentation; non-interference algorithm

1　引　言

验证码的全称为全自动区分计算机和人类的图灵测试(Completely Automated Public Turing test to tell Computers and Humans Apart,CAPTCHA),由卡内基梅隆大学研究人员于2000年提出[1].在日益增多的互联网活动中验证码起着重要作用,诸如在网站注册、登录和密码找回等环节中能有效地抵御自动化攻击[2].按照信息类型的不同,现有的验证码可分为两类:一类是基于视觉的验证码,包括文本类验证码和图像类验证码;另一类是基于听觉的语音类验证码[3].图像类验证码通过识别验证码中特定对象所属的类别[4]实现验证,语音类验证码则采用匹配语音中包含的验证信息实现验证[5-6].虽然这两种新兴的验证码丰富了验证方式,但是在对交互时

间和安全性要求较高的场景中，这两种验证码都呈现出适应性受限、交互友好性不足和安全性不佳等问题[7]. 相反，由于文本类验证码具有交互方式简单、密码空间大且场景适应性强等特点[8]，在实际应用中被广泛接受. 调查发现，在 Alexa 发布的全球综合排名前 50 的网站中，80％的网站在登陆、注册、多次输错密码或者密码找回环节中使用验证码来抵御自动化攻击，其中包含微软、百度、ebay 等在内的 60％的网站均在使用文本类验证码[9]. 因此，深入研究文本类验证码的安全性对于改进传统验证码生成方法具有重要意义. 本文研究对象及下文所提到的验证码皆为文本类验证码.

按照识别过程中是否需要对字符进行分割，文本类验证码识别方法可分为分割识别和整体识别两种[10]. 分割识别是目前比较普遍的验证码破解方法，Chellapilla 等人的研究[11]表明，对验证码中字符的有效分割可大大提高识别准确率. 早期验证码识别技术以分割识别为主，这一时期的验证码以 Captchaservice.org[12]网站上的验证码为代表. 其特点为背景干扰信息较少或者几乎没有，字符也缺少扭曲、旋转及粘连等复杂的变换，防御效果有限. Yan 等人[12]使用像素统计的方法对这类验证码实现了破解. 此后验证码设计者改进了生成算法，加入背景干扰信息，其以微软[13]早期的验证码为典型代表. 而 Yan 等人[13]在 2008 年使用投影算法对其进行了准确率高达 90％的有效分割，破解成功率达 60％. 在连续两轮攻击与防御的对抗后，为了更好的抵御分割算法，设计者对验证码进行了进一步的改进，增加了字符扭曲、旋转和粘连等多种复杂变换，同时对背景干扰信息也进行更为复杂的变换[14-15]. 针对此类验证码，Li 等人[16]通过改进的滴水算法[17]对轻微粘连的验证码进行破解，其分割准确率达到 78％. 同时，Gao 等人[8]也通过 Gabor 滤波提取字符笔画，利用图搜索查找最优字符笔画组合的方式进行字符分割，对 reCAPTCHA 的破解准确率达到 77％. 对于字符无法从背景干扰信息中剥离并逐一分割的验证码，有研究者采取整体识别的方案[18-19]，Wang[20]采用基于形状上下文整体识别的算法来破解早期天涯论坛验证码，其准确率为 27.7％. Yi 等人[10]使用一种对特征点进行匹配分析的方法对 2014 年美团网验证码整体识别，准确率达到 88％.

目前，文本验证码的安全性主要靠背景干扰信息和字符粘连两个因素来进行保证[15]，这两个安全特征都在不同程度上增加了识别难度和分割难度[21]. 背景干扰信息让字符特征更为模糊，在增加机器分析难度的同时也让分割更加困难；字符粘连在增加字符分割难度的同时也扰乱了字符的特征信息.

但是，仅仅依靠这两个方面的强化是否能够完全保证文本验证码的安全性呢？本文对这一问题进行了探讨. 具体而言：本文利用 CGAN 在计算机视觉领域图片生成和图片风格转换方面的优势[22-23]，在预处理阶段两次使用 CGAN 来生成新的验证码. 第一次是根据真实验证码生成无干扰信息的验证码，第二次是对无干扰验证码中字符间距进行拉伸，从而为分割阶段提供“背景更干净、字符间距更大”的验证码. 通过优化组合的分割算法对预处理后的验证码进行分割，再把分割之后的单个字符通过 GoogLeNet 进行识别.

实验结果表明，本文方法对于含有背景干扰信息且字符扭曲、旋转和粘连的验证码有很好的识别效果，对某电商平台目前正在使用的验证码识别准确率达到 86％. 本文分割算法与传统分割算法相比，其有效性不会受到验证码字符大小、样式、扭曲程度、旋转角度、空心字体等因素影响，具有很好的普适性. 实验还发现本文方法对于空心字符填充效果优于传统的空心字符填充算法.

2 理论基础

本文系统在预处理阶段和字符识别阶段分别使用了 CGAN 和 GoogLeNet，本节将对其理论基础进行简要介绍.

2.1 卷积神经网络(CNN)与 GoogLeNet

GoogLeNet 是当前最经典的 CNN 之一，CNN 通过共享权值、局部连接和下采样过程不仅有效地解决了全连接网络所面临的参数过多、无空间结构和无法深入的问题，还提高了网络训练速度、模型准确率及鲁棒性[24]. GoogLeNet 则是 CNN 朝着更深和更宽方向探索的代表，其中卷积层运算的过程如式(1)[25]所示：

$$x_j^l = f\left(\sum_{i\in M_j} x_j^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

在式(1)中，x_j^l表示求得的第 l 层的第 j 个特征图，$f(\cdot)$表示本层所使用的激活函数，k 表示卷积核，M_j表示输入的特征图，“$*$”是矩阵运算中的卷积乘法，b_j表示添加的偏置值. GoogLeNet 中激活函数使用的是 ReLU，其数学描述如式(2)所示：

$$f(x) = \max(0, x) \tag{2}$$

ReLU 相比于其他的激活函数有着很大的计算优势，可以让网络的设计更加深入. 池化层的主要任务是用来进行降采样，降采样是指提出特征图中关键的信息来减少参数数量. 常用的降采样方法有最大值采样法和平均采样法，其过程可描述为式(3)所示：

$$x_j^l = f(\beta_j^i down(x_j^{l-1}) + b_j^l) \tag{3}$$

上文式(3)中 x_j^l 为求得的 l 层的第 j 个特征图，其中 $down(\cdot)$ 指选取的降采样函数，β 是加权系数，b_j 为偏置值.

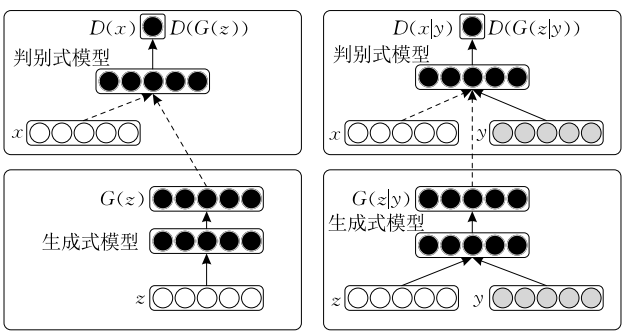
卷积神经网络的训练比全连接神经网络复杂，但基本原理是相似的. 过程为：通过链式求导得出损失函数对于每个权重的偏置值，再根据梯度下降公式和学习率来逐步更新权重.

2.2　生成式对抗网络(GAN)与条件生成式对抗网络(CGAN)

本文系统两次使用 CGAN 来生成新的验证码，CGAN 在继承 GAN 基本框架和优化方式的基础上对于图片生成方式做出了改进，本节将探究两种网络的基本框架和优化方式.

2.2.1　生成式对抗网络(GAN)

GAN 的诞生是受到博弈论的启发，其中博弈双方分别是生成式模型和判别式模型[22]. 其网络训练流程结构如图 1(a)中所示. 图中可微分函数 G 和 D 分别表示生成式模型和判别式模型，z 和 x 分别表示随机变量和真实数据. $G(z)$ 是指生成式模型生成的数据，通过判别式模型之后输出概率为 $D(G(z))$. 真实数据 x 通过判别式模型后输出概率为 $D(x)$.



(a) 生成式对抗网络(GAN)　　(b) 条件生成式对抗网络(CGAN)

图 1　GAN 与 CGAN 训练流程图

用于图片生成任务的 GAN，其生成式模型和判别式模型多选用定制的 CNN 实现. GAN 网络的训练机制复杂于传统的神经网络，在优化判别式模型时先假设生成式模型为固定的可微分函数，其训练过程类似于最小化交差熵的过程，目标函数可描述为式(4)：

$$Obj^D(\theta_D,\theta_G) = -\frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] - \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{4}$$

式(4)中 z 和 x 分别表示采样于特定分布 $p_z(z)$ 的随机变量和采样于真实分布的真实样本，$\mathbb{E}(\cdot)$ 表示计算期望值. 判别式模型优化过程需要考虑 $G(z)$ 和 x 两种输入影响，当判别式模型的输入为 $G(z)$ 时，判别式模型需要使得输出概率 $D(G(z))$ 趋近于 0，而生成式模型则需要使得输出概率 $D(G(z))$ 趋近于 1. 当输入为真实数据 x 时，判别式模型只需要使得 $D(x)$ 趋近于 1 即可.

在固定判别式模型时，生成式模型的目标函数可描述为：$Obj^G(\theta_G) = -Obj^D(\theta_D,\theta_G)$，因此 GAN 的优化问题可视为一个极小极大值问题. GAN 整体的目标函数可描述为

$$\mathcal{L}_{\text{GAN}}(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{5}$$

GAN 的训练过程可总结为：先通过固定生成式模型来提高判别式模型对于 $D(x)$ 和 $D(G(z))$ 的灵敏度，再固定判别式模型来提高生成式模型生成数据 $G(z)$ 的有效性. 当且仅当训练好的判别式模型无法区分生成数据时训练终止.

2.2.2　条件生成式对抗网络(CGAN)

传统的 GAN 根据随机噪声 z 来生成图片的方式太过自由，缺乏用户控制能力. Mirza 等人通过在生成式模型和判别式模型中加入约束条件来解决该问题，并提出了新的网络 CGAN. 其训练流程结构如图 1(b)中所示，相比于传统 GAN，CGAN 在 G 和 D 中分别加入约束条件 y. 图 1(b)中 $G(z|y)$ 是指生成式模型在约束条件的约束下生成的数据，和约束条件同时输入判别式模型之后输出条件概率为 $D(G(z|y))$. 真实数据 x 和约束条件同时输入判别式模型后输出条件概率为 $D(x|y)$.

CGAN 的训练过程和 GAN 的训练过程相同，依次固定生成式模型来优化判别式模型，固定判别式模型来优化生成式模型. 其优化问题可视为一个带条件概率的最小最大值问题，其整体目标函数可描述为

$$\mathcal{L}_{\text{cGAN}}(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z|y)))] \tag{6}$$

优化过程中当判别式模型的输入为真实数据 x 和约束条件 y 时，判别式模型需使得条件概率

$D(x|y)$趋近于1.当判别式模型的输入为生成数据$G(z|y)$和约束条件y时,判别式模型需使得条件概率$D(G(z|y))$趋近于0,此时生成式模型则需要使得条件概率$D(G(z|y))$趋近于1,来提高生成效果.

3 系统设计与实现

上节中对系统需要的理论基础进行了分析,本节则详细介绍如何将这些理论基础用于解决验证码识别问题.

3.1 系统框架

如图2系统框架图所示,本文识别系统将验证码处理过程分为预处理、字符分割、字符识别三个阶段共六个步骤.点划线下部分为百度验证码每个环节处理之后的效果图.

图2中系统输入为真实的百度验证码,在保证原图长宽比的同时将其扩大到统一尺寸并转换为灰度图像,灰度化处理可减少颜色对于网络的影响.对于含有背景干扰信息的验证码,本文提出了一种高效且普遍适用的去干扰网络(Remove Background Interference Network,RBNet).训练好的去干扰网络(RBNet)能够在含干扰信息验证码的约束下生成无干扰信息的验证码.去除干扰信息之后则通过字符间距拉伸网络(Character Spacing Stretch Network,CSNet)对验证码中的字符间距进行拉伸,以此来优化分割过程.字符分割阶段将CFS(Color Filling Segmentation)算法[13]、投影算法[13]和滴水算法[17]优化组合起来对预处理后的验证码进行分割.系统首先利用CFS算法对经过字符间距拉伸的验证码进行分割,对于CFS算法无法分割的部分(即通过CSNet拉伸后依然存在轻微粘连的部分),通过投影算法预估字符粘连位置(如图2分割阶段所示),再以此位置为起始点采用滴水算法进行分割.在得到高效分割的单个字符之后,最后利用识别网络(GoogLeNet)对字符进行识别,根据验证码字符个数组合识别结果,最后输出验证码文本内容.

本文系统中RBNet和CSNet的实现皆基于CGAN,两者都使用程序模拟的验证码作为样本进行训练.训练RBNet时,其约束条件为含有干扰信息的验证码,监督标签为无干扰信息的验证码,生成式模型在约束条件的约束下生成无干扰信息的验证码.CSNet则是以字符粘连的验证码作为约束条件,以不含粘连字符的验证码作为监督标签,通过生成式模型来生成字符间距较大的验证码.由于CSNet对字符间距的有效拉伸,基本已经消除了字符粘连,遂使用对无字符粘连更有效的CFS算法先进行分割,粘连部分则使用对字符轻微粘连更有效的滴水算法进行分割.相对于传统CNN,GoogLeNet有着更高的识别效率和更短的训练时间,最后环节选用GooLeNet对分割的单个字符进行有效识别.

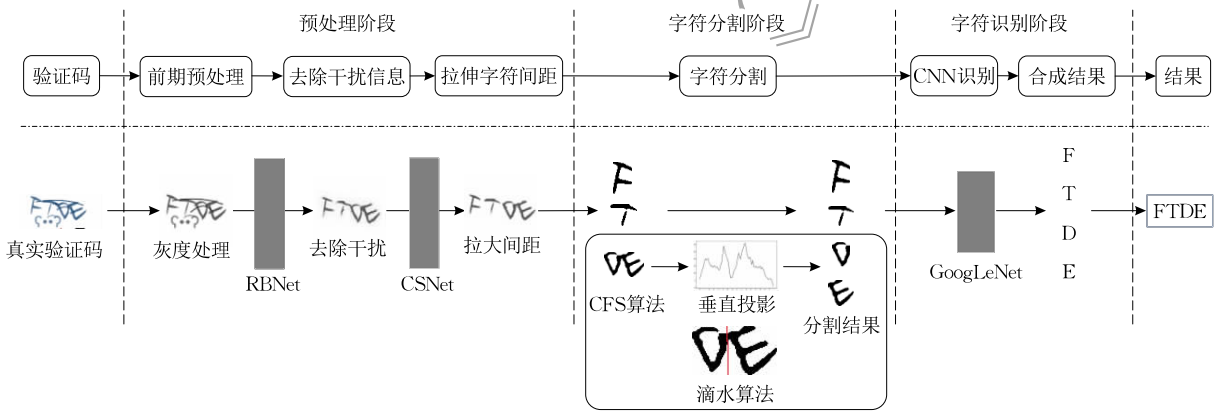


图2 系统框架图

3.2 预处理阶段

在上述系统的具体实现过程中,我们发现存在真实样本难以获取的问题.因此,本文设计了程序模拟验证码的生成过程.在获得大量训练样本之后,即可完成预处理阶段两种网络的训练.完成训练后,预处理阶段便可以通过前期预处理、去除干扰信息和拉伸字符间距三项操作为字符分割阶段提供"背景更干净、字符间距更大"的验证码.

3.2.1 模拟验证码生成

对于神经网络而言,训练样本数量和网络性能呈现较强的相关性.为了增强网络的有效性,在网络训练时需要大量的验证码作为训练样本.由于:

(1)现在大多数网站都做了防爬虫处理,难以通过自动化程序获取大量真实的验证码.

(2)即使通过自动化程序获取到了部分验证码，也由于没有对应的标签，而需要大量人工标注工作.

(3)RBNet 与 CSNet 都需要进行监督训练，RBNet 需要以不含干扰信息的验证码作为监督标签，CSNet 需要以无字符粘连的验证码作为监督标签，自动化程序只能获得真实验证码.

结合以上因素，本文通过让程序模拟真实验证码的分布特征，来生成模拟验证码作为训练样本.

通过对真实验证码进行分析，本文归纳总结出了 11 个特征，使得任何验证码均可通过图 3 中的 11 种特征进行描述. 其中与干扰信息相关的特征有两种，与字符相关的特征有九种，图 3 中括号中的值表示从该区间随机取值.

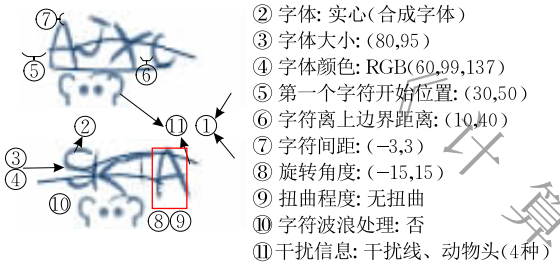
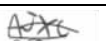
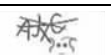


图 3　真实验证码特征分析图

表 1 呈现了真实验证码和三种由程序模拟生成供训练使用的验证码，其中包括：完整模拟出真实验证码的所有特征，未做任何变换的完整模拟验证码；无干扰信息验证码是在完整模拟验证码的基础上去除了干扰信息；无字符粘连验证码是在完整模拟验证码的基础上去除了干扰信息并增加了字符间距，保证无字符粘连.

表 1　程序模拟验证码效果展示表

验证码	真实验证码	程序模拟验证码		
		完整模拟验证码	无干扰信息验证码	无字符粘连验证码
京东			W5EU	W5EU
百度			AJXC	AJXC
新浪			m8eFy	m8eFy

完整模拟验证码和无干扰信息验证码除是否含有干扰信息外，其他特征均相同. 无干扰信息验证码和无字符粘连验证码除字符间距外，其他特征均相同. 在 RBNet 的训练过程中，将完整模拟验证码作为训练样本(即 CGAN 中的约束条件)，将无干扰信息验证码作为监督标签. 在 CSNet 的训练过程中，将无干扰信息验证码作为训练样本，将无字符粘连验证码作为监督标签.

3.2.2　前期预处理

前期预处理包含两项操作：统一尺寸和灰度化处理. 统一验证码尺寸是因为 RBNet 和 CSNet 需要验证码长宽作为超参数进行训练. 为了不让验证码发生形变，通过等比例扩大或缩小至长边达到预设值后，对短边两侧同时使用白色填充的方法进行扩大或缩小(效果如图 4 所示). 灰度化处理可减少颜色对于网络的影响，提高网络的鲁棒性.

(a) 真实验证码　　(b) 统一尺寸　　(c) 灰度化处理

图 4　前期预处理示意图

3.2.3　去除干扰信息(RBNet)

由于复杂的背景干扰信息不仅可以模糊和扰乱特征信息，还会影响分割准确率，所以有效去除干扰信息是保证识别准确率的前提条件. 传统去干扰法通常是根据不同干扰信息的特点，设计针对性的去除算法[26-28]. 当验证码设计者将多种类别的干扰信息随机组合使用时，传统的去除算法便难以奏效.

本文去干扰法的设计抛弃了传统的设计思想，通过神经网络 RBNet 来生成新的去除干扰信息的验证码，从而避免了为每种干扰信息设计新的去噪算法.

RBNet 中生成式模型使用的是 U-Net 网络[29]，判别式模型使用的是 PatchGAN[30] 分类器. U-Net 是一种用于图片生成的全卷积结构神经网络，其结构如图 5 中生成式模型所示. 与普通编码器(Encoder-Decoder)网络的不同点在于 U-Net 为解码前后对应的特征图建立了连接. 普通编码器是通过将采样降到更低的维度，然后再将采样升到原始的尺寸，而 U-Net 将编码阶段和解码阶段对应特征图通过通道进行了拼接，这样保存了不同分辨率下像素级别的细节内容，使生成式模型生成的验证码和监督标签(表 1 中无干扰信息验证码)在细节上更加相似. 传统 GAN 中判别式模型是将需要判别的图片映射到一个矢量，以此判定生成式模型的生成效果. PatchGAN 用于 CGAN 时，将需要判别的图片(生成器生成的验证码)和监督标签(表 1 中无干扰信息验证码)对应分为多个大小为 $N\times N$ 的块. 不同的块之间相互独立，每次只需要输入一个对应块，通过卷积操作得到该块的生成效果. 将所有块的输出结果合并到一个特征图，最后将该特征图取平均值作为最终判别式模型的输出. 这样可让输入图片的维度降低，计算需要的时间和参数量减少，并且可以计算任

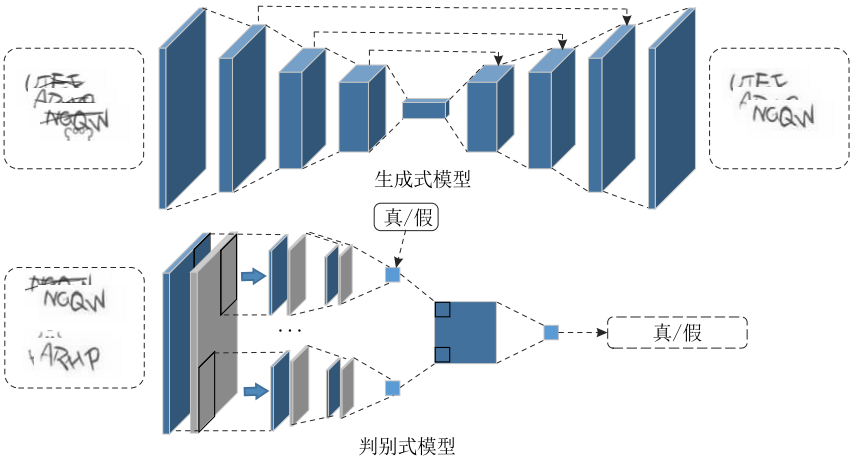


图 5 RBNet 网络结构图

意大小的图片.

传统用于生成图片的 GAN 中生成式模型 G 学习一个从随机噪声 z 到真实数据 x 的映射,可表达为 $G:z\rightarrow x$. RBNet 在用于验证码生成时,生成式模型 G 学习一个从含干扰信息的验证码 y 和随机噪声 z 到无干扰信息验证码 x 的映射,可表达为 $G:\{z,y\}\rightarrow x$,网络的目标函数如上文式(6)所示.

生成式模型 G 通过不断学习来最小化目标函数,而相反的判别式模型 D 则通过不断学习来最大化目标函数,优化函数如下:

$$G^{*}=\arg\min_{G}\max_{D}\mathcal{L}_{\mathrm{cGAN}}(G,D) \tag{7}$$

研究[31-32]表明,若在 CGAN 的优化过程中令其原始目标函数加入传统的损失函数,生成式模型会有更好的效果. 加入传统损失函数之后,判别式模型的功能不变,生成式模型不仅要生成去除干扰的验证码,还要求生成的验证码无限的逼近于监督标签(表 1 中无干扰信息验证码). 常用 L1 distance L_1: $d_{ii'}(1)=|x_{i1}-x_{i'1}|+|x_{i2}-x_{i'2}|$ 和 L2 distance L_2: $d_{ii'}(2)=((|x_{i1}-x_{i'1}|^2+|x_{i2}-x_{i'2}|^2)^{1/2})$ 来表示两者无限逼近时的距离. 相比于 L1 distance,L2 distance 的模糊度更大. 本文此处使用 L1 distance,目标函数可表示为

$$\mathcal{L}_{\mathrm{L1}}(G)=\mathbb{E}_{x,y\sim p_{\mathrm{data}}(x,y),z\sim p_z(z)}[\|y-G(z|y)\|_1] \tag{8}$$

图 6 是 RBNet 的优化过程,其中约束条件 y 是指含有干扰信息的验证码(表 1 中完整模拟验证码),目标 x 是指监督训练的监督标签(表 1 中无干扰信息验证码). 图 6 中(a)是判别式模型的预训练过程. 图 6 中(b)是生成式模型的训练过程,随机噪声 z 在约束条件 y 的约束下生成无干扰信息的生成验证码. 图 6 中(b)和(c)共同构成了对抗训练过程,固定判别式模型,将生成验证码和约束条件 y 同时送入判别式模型中,根据判别结果来指导生成式模型进行优化. 固定生成式模型,将生成验证码和约束条件 y 同时送入判别式模型中,让判别式模型更敏锐地区分生成验证码和监督标签的不同. 如此反复,直到训练好的判别式模型无法区分生成验证码的真假时训练完成.

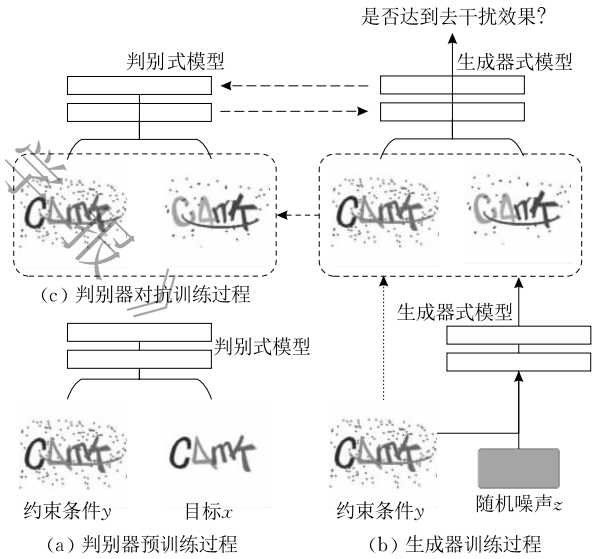


图 6 RBNet 训练流程图

在 RBNet 的构建中,对于加入的传统损失函数,本文分别尝试 L1 distance 和 L2 distance,其结果对比图如图 7 所示. 实验表明,两者均可以满足去除干扰信息的需要,L1 distance 可以保存更多的特征信息,

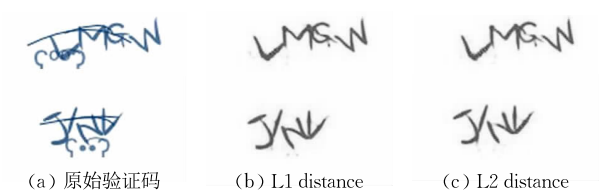


图 7 RBNet 中使用 L1 和 L2 去干扰效果对比图

效果更好. 在使用 L1 distance 时优化公式如式(9)所示.

$$G^*=\arg\min_G\max_D\mathcal{L}_{cGAN}(G,D)+\lambda\mathcal{L}_{L1}(G) \quad (9)$$

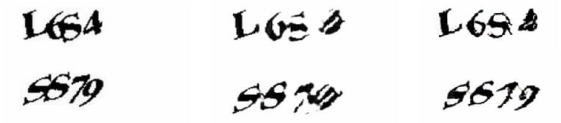
3.2.4　拉伸字符间距(CSNet)

有研究表明[11,14],字符粘连是保障验证码安全最有效的方式. 粘连字符难以分割的关键就在于其空间分布和粘连方式,本文先通过 RBNet 去除干扰信息再通过 CSNet 拉伸字符间距,当字符间距变大之后使用组合优化的分割算法进行分割. 这样可显著提高分割准确率.

CSNet 的实现也是基于 CGAN,其网络结构与 RBNet 基本相似,如图 5 所示. RBNet 的应用场景相当于验证码风格的转换,而在 CSNet 中则涉及字符轮廓轻微变化. 鉴于 U-Net 网络编码和解码时在同级特征图之间建立连接的特殊网络结构,使其对于生成轮廓轻微变化的图片依然有效. 本文通过大量实验也说明了,CSNet 能够根据字符粘连的验证码生成无字符粘连的验证码有很好的效果. PatchGAN 优化的计算方式,在提高了准确率的同时大量减少了训练过程. 本文 CSNet 的实现过程中生成式模型和判别式模型依然选用的是 U-Net 和 PatchGAN.

在验证码间距拉伸任务中,生成式模型 G 学习一个从字符粘连验证码 y 和随机噪声 z 到无字符粘连的验证码 x 的映射,可表达为 $G:\{z,y\}\rightarrow x$,网络的目标函数与上文式(6)相同. 生成式模型优化时在其原始目标函数中依然加入传统的损失函数,和 RBNet 不同之处在于 CSNet 使用 L2 distance 来表示两者无限逼近时的距离. 由于 L2 distance 拥有模糊度更大的特征,对于字符轮廓上有轻微变化的验证码有更好的生成效果,本文实验部分也说明了这一点(新浪微博验证码的拉伸效果如图 8 所示). 网络的目标函数如式(10)所示:

$$\mathcal{L}_{L2}(G)=\mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[\|y-G(x,z)\|_2] \quad (10)$$



(a) 原始验证码　(b) L1 distance　(c) L2 distance

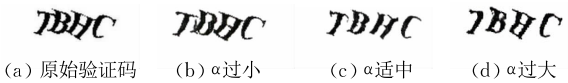
图 8　CSNet 中使用 L1 和 L2 拉伸效果对比图

在网络的训练时采用交替优化,其优化函数如式(11)所示:

$$G^*=\arg\min_G\max_D\mathcal{L}_{cGAN}(G,D)+\lambda\mathcal{L}_{L2}(G) \quad (11)$$

CSNet 的训练过程与图 6 RBNet 的训练流程图基本相同,约束条件 y 是指字符粘连且无干扰信息的验证码(表 1 中无干扰信息验证码),目标 x 是指监督训练的监督标签(表 1 中无字符粘连验证码). 通过不断的对抗训练,直到训练好的判别式模型无法区分生成验证码的真假则认为训练完成.

在 CSNet 训练时,监督标签(表 1 中无字符粘连验证码)相对于约束条件 y(表 1 中无干扰信息验证码)字符间距统一增加了 α. 图 9 中展示了新浪微博验证码不同 α 值的监督标签所生成的不同效果的验证码. 由此也说明不同的 α 值直接影响着 CSNet 的生成效果. 本文选择 α 值所使用的策略是:先充分模拟真实验证码的各种特征(字符旋转角度、扭曲程度、间距等)的随机分布,同步生成约束条件 y 和监督标签. 在生成监督标签时,字符间距较约束条件 y 统一增加 α,保证当所有验证码中不存在字符粘连时取最小的 α 值.



(a) 原始验证码　(b) α过小　(c) α适中　(d) α过大

图 9　新浪微博验证码不同 α 生成效果对比图

3.3　字符分割阶段

由于本文预处理阶段 CSNet 对验证码字符的间距进行了拉伸,有效消除字符粘连. 因此,分割阶段将适用于无字符粘连的 CFS 算法作为主要分割算法,将适用于字符轻微粘连的滴水算法作为辅助分割算法. 分割过程使用的分割算法如算法 1 所示,处理流程如图 10 所示,对预处理之后的验证码先采用 CFS 算法进行分割. 统计单个字符的最大长度和最小长度,对分割结果进行分析,筛选出粘连部分. 根据统计值确定粘连字符个数,并用滴水算法进行分割. 滴水算法在分割前先对粘连部分进行垂直投影,按粘连的字符个数进行等分. 在每个等分点左右三分之一领域内找像素直方图的极小值点作为滴水算法的起始位置,以此点进行滴水分割.

算法 1. CharacterSegmentation(*ProcessedImage*).

输入:*ProcessedImage* 预处理之后的图片

输出:*M* 分割结果的集合

1. *M*←用来存放分割结果的集合
2. *N*←*CFS*(*ProcessedImage*)
3. FOR *each* IN *N*:
4. 　*characterNumber*←分析分割结果中字符个数(*each*,*MAXCharLen*,*MINCharLen*)
5. 　IF *characterNumber*==1:
6. 　　*M.add*(*each*)
7. 　ELSE:

8. *startPoint*←垂直投影(*each*)
9. *S*←滴水分割(*each*,*startPoint*)
10. *M.add*(*S*)

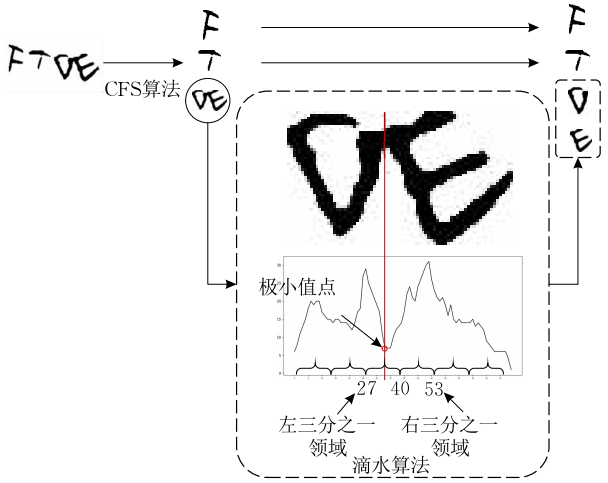


图 10 本文分割算法处理流程图

3.4 字符识别阶段

通过前两个阶段的工作,在此已经得到了有效的分割的单个字符,本阶段需要通过神经网络对单个字符进行识别.对于图片识别任务,多使用 CNN 来完成.三种经典 CNN 网络分别是:LeNet5[24],AlexNet[33]和 GoogLeNet[34-35],本节将通过多个不同角度对识别网络效果进行探究.

本文在此处使用百度、京东和支付宝三个网站真实验证码分割之后的单个字符作为测试样本,使用 CSNet 的监督标签(表 1 中无粘连字符验证码)分割之后的单个字符作为训练样本,分别对 LeNet5、AlexNet 和 GoogLeNet 识别效果进行评估.鉴于 LeNet5 结构过于简单,本文对其进行了优化,主要包括将其网络层数由七层深化为十层,将输入图片的尺寸由 32 改为 128,并借鉴了 AlexNet 中的非线性激活函数和防过拟合方法.AlexNet 和 GoogLeNet 使用经典网络结构.表 2 呈现的是三种网络字符识别准确率对比表,实验结果表明:使用 GoogLeNet 的准确率优于 AlexNet 和优化过的 LeNet5.使用同样的数据集大小并迭代训练同样的次数,优化过的 LeNet5 需要的时间最短,其次是 GoogLeNet,而 AlexNet 则需要的时间最长,鉴于准确率和训练效率两个因素,本文在之后的实验中都使用GoogLeNet 作为识别网络.

表 2 三种网络字符识别准确率对比表

验证码	LeNet5(优化后)/%	AlexNet/%	GoogLeNet/%
百度	57.11	59.15	59.50
京东	95.35	96.98	97.11
支付宝	77.34	78.65	78.88

识别网络在训练过程中,训练数据集的大小和迭代次数都会对测试集准确率有所影响.本文使用 GoogLeNet 对训练集大小和迭代训练次数对测试集准确率的影响做了评估,表 3 是五种不同大小训练集对应的测试集准确率.可以看出训练集数量从 5000 到 25000 测试集的准确率都有较大的提升,而从 25000 到 35000 提升的速度变缓,百度在 30000 时还出现了更高的准确率.鉴于准确率和训练效率的影响,本文在识别阶段选用的训练集大小为 30000.从图 11 中可以看出京东验证码训练集准确率和 Loss 值很快就达到了最优值,而测试集的准确率和 Loss 值则是在 100 次迭代时达到最优值.为保证实验的准确性,本文之后的实验中识别网络训练迭代次数都设为 200.

表 3 五种不同大小训练集字符识别准确率对比表

验证码	百度%	360/%
5000	52.64	74.32
15000	58.54	83.43
25000	59.32	89.17
30000	59.50	89.20
35000	59.34	89.20

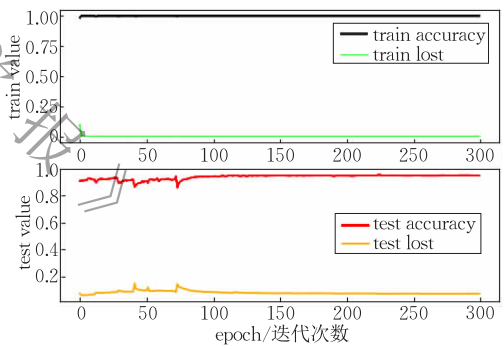


图 11 京东验证码训练效果图

4 实验结果与分析

本节将对本文预处理效果和验证码识别效果分别进行实验评估,并对实验结果进行详细的分析.

4.1 预处理效果评估

本节介绍了本文实验中所用到的两个数据集,通过效果图对比的方式展示 RBNet 的去干扰效果,通过分割准确率对比的方式展示 CSNet 的拉伸效果.

4.1.1 实验数据集

本文实验数据均来自 Alexa 全球综合排名前五十的网站中三十个使用文本类验证码的网站,数据集采集于 2017 年 12 月.实验分为两个数据集,每种

验证码包含 200 张. 数据集 1 如图 12 所示，是从以上三十个网站中筛选出的三种不同粘连程度的验证码. 该数据集主要用于展示 CSNet 对于不同粘连程度的验证码的拉伸效果. 数据集 2 如图 13 所示，把以上三十个网站分为电商类、社交类、搜索类、知识类四类，剔除相似类型验证码之后，从每个类别中取出典型的几种. 个别网站验证码若含有多种风格则用其中的一种进行分析，个别验证码中若含有多种字符个数则选取比例最高的字符个数进行实验.

图 12　数据集 1 三种不同粘连程度的验证码

图 13　数据集 2 验证码及分类

4.1.2　RBNet 去干扰效果评估

在数据集 2 中百度和搜狐验证码为含有干扰线的验证码，图 14 中(a)展示的是通过形态学方法去除干扰信息的效果图. 由于这两种验证码在设计时将干扰线粗细程度设计的与字符笔画粗细程度相同，导致使用形态学去除法时更换多种不同的核都无法去除干扰信息. 除此之外，这两种验证码中还添加有其他类别干扰信息，所以单纯的使用形态学的方法根本无法将干扰信息有效去除. 图 14 中(b)展示了 RBNet 的去干扰效果，可以看到这两种验证码中的干扰信息都被完整的去除，因此可以说明 RBNet 的效果优于传统的去干扰法.

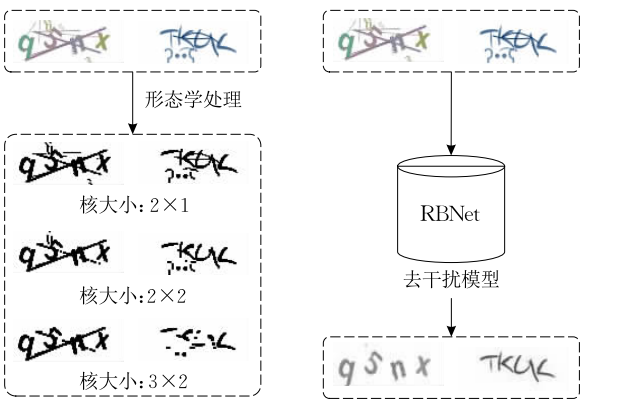


图 14　去除干扰线效果对比图

RBNet 不仅可以高效地去除干扰信息，对空心字符在去除干扰信息的同时还能进行高效的填充. 图 15 中(a)展示了传统字符填充算法的效果图，该算法的实现参考了文献[15]. 其实现主要是利用了图像学知识，填充效果会受到验证码中干扰线的影响，对部分字符无法实现有效填充. 图 15 中(b)展示了 RBNet 对空心字符的填充效果，可以看出其不仅可以有效的去除干扰信息还能高效的对空心字符进行填充.

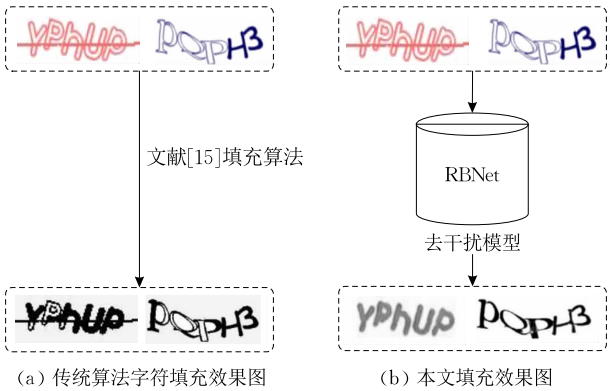


图 15　空心字符填充效果对比图

为了验证本文去干扰方法的普适性，图 16 展示了数据集 2 中所有含有干扰信息验证码的去干扰效果. 可以看出现存验证码中干扰信息多为不同类别干扰信息的组合，而传统的去干扰法难以对此类组合干扰信息有效去除. 在使用本文提出的 RBNet 去干扰过程中，则不用考虑干扰信息的分类和组合，因此具有很好的有效性和普适性.

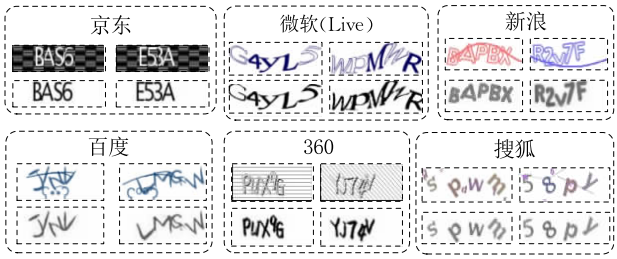


图 16　数据集 2 中验证码去干扰效果展示图

4.1.3　CSNet 字符拉伸效果评估

通过 CSNet 对字符间距的有效拉伸，为分割阶段提供了极大的优势. 本节将通过对比拉伸前后分割准确率的变化来评估 CSNet 的有效性.

(1) 分割准确率评估标准

分割准确率的评估标准是观察分割之后字符的完整性. 若字符轮廓完整，没有出现缺失也没有增加

多余部分，则认为是正确分割的字符. 图 17 中第二个验证码由于字符 S 有多余部分，则认为字符分割出错. 分割实验分别统计字符分割准确率(式(12))和验证码分割准确率(式(13))，一个验证码中的字符全部分割正确，则认为这个验证码分割正确.

$$\text{字符分割准确率}=\frac{\text{正确分割字符个数}}{\text{字符总数}}\times 100\% \tag{12}$$

$$\text{验证码分割准确率}=\frac{\text{正确分割验证码个数}}{\text{验证码总数}}\times 100\% \tag{13}$$

(a) 原始图片 (b) 分割结果

图 17 验证码分割效果展示图

(2) 分割效果分析

本实验中将对五种分割过程进行效果评估，分别是投影算法、CFS 算法、滴水算法、本文分割算法和 CSNet 拉伸后的本文分割算法. 前四个分割过程在分割之前皆采用 RBNet 去除干扰信息，最后一种分割过程在采用 RBNet 去除干扰信息之后通过 CSNet 拉伸字符间距再进行分割.

从图 18 和图 19 三种不同粘连程度的验证码分割情况中可以看出，本文优化组合的分割算法在字符分割准确率和验证码分割准确率上均优于或等于其他三种分割算法. 而通过 CSNet 拉伸之后，字符分割准确率和验证码分割准确率皆有明显的提升，且均高于其他算法的分割效果. 通过其他三种分割算法和本文分割算法效果对比，说明了本文分割算法的高效性. 通过是否采用 CSNet 对字符间距进行拉伸效果对比，说明了 CSNet 的高效性.

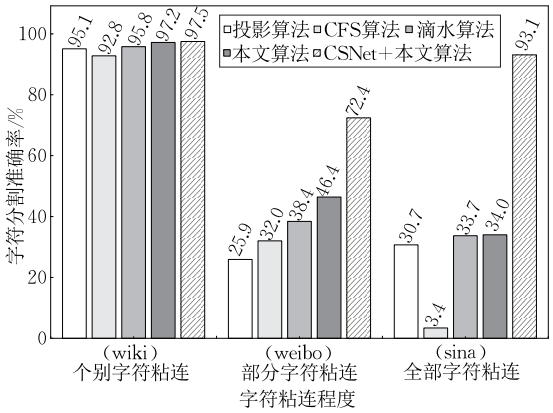


图 18 字符分割准确率对比图

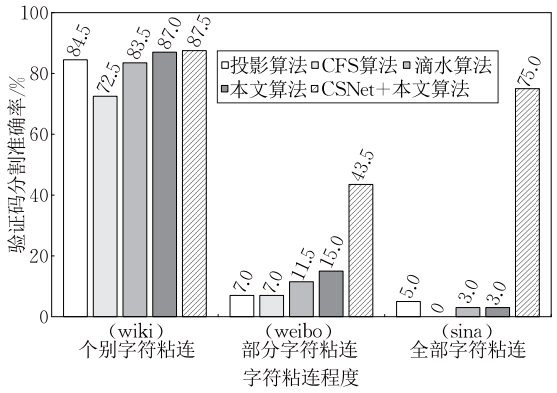


图 19 验证码分割准确率对比图

对于个别字符粘连的维基百科验证码，几种算法效果差距较小. 对于微博验证码，其他三种算法字符分割准确率和验证码分割准确率最高分别是 38.4% 和 11.5%，采用本文分割算法在未通过 CSNet 拉伸之前准确率即可提升至 46.4% 和 15%，通过 CSNet 拉伸之后准确率则提升至 72.4% 和 43.5%. 对于新浪验证码，CSNet 拉伸之前本文分割算法和滴水算法差距较小，通过 CSNet 拉伸之后则有着明显的提升，图 20 为本文系统处理新浪验证码时各个阶段的效果图.

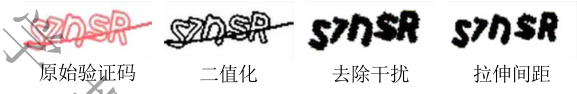


图 20 新浪验证码处理效果图

表 4 呈现了数据集 2 中所有网站验证码字符分割准确率和验证码分割准确率的对比表. 其中字符完全粘连的京东验证码分割准确率最高，由于该验证码中字符未进行扭曲旋转变换并采用了单一的背景作为干扰信息. 该背景干扰可通过 RBNet 清晰去除，未扭曲旋转的字符，通过 CSNet 可高效的拉伸字符间距. 实验表明本文方法对于空心字符验证码

表 4 数据集 2 字符分割准确率验证码分割准确率展示表

分类	网站名称	字符分割准确率/%	验证码分割准确率/%
电商类	京东	97.20	90.5
	支付宝	78.88	52.0
	易贝	66.50	30.5
社交类	微软	70.70	36.0
	微博	72.40	43.5
搜索类	百度	61.00	23.5
	360	89.20	59.5
知识类	维基百科	97.50	87.5
	搜狐	95.13	83.0
	新浪	93.10	75.0

有更好的效果，这十种验证码中京东、新浪和 360 三个网站的验证码中都使用了空心字符，其验证码分割准确率分别为 90.5%、75%和 59.5%.

4.2　识别效果评估

(1)实验评估标准

本节实验将对本文系统识别准确率进行评估.实验数据集选用上文中两个数据集，统计结果时一个字符识别出错则认为此验证码识别错误，结果忽略字符大小写的区别，识别准确率计算公式如式(14)所示：

$$\text{验证码识别准确率}=\frac{\text{正确识别验证码个数}}{\text{验证码总数}}\times 100\% \tag{14}$$

在充分的对比优化后的 LeNet5、AlexNet 和 GoogLeNet 的识别效果后，字符识别阶段选择 GoogLeNet 来对分割之后的单个字符进行识别.训练时每次采用 64 张图片同时进行参数更新和反向传导，将整个训练数据集迭代训练 200 次，每训练完一次对数据集进行洗牌操作.

(2)识别结果与分析

从图 21 分割准确率和识别准确率对比图可以看出识别准确率会受到分割准确率的影响，而识别准确率通常都会低于分割准确率.这部分准确率损失分别来自于：①程序模拟验证码和真实验证码字体差异；②识别网络固有的设计缺陷.其中维基百科在训练时没有找到和真实验证码相同的字体，使用一种相似的字体进行训练，验证码分割准确率和识别准确率相差 9%.微博和新浪验证码在训练时使用相同字体，其验证码分割准确率和识别准确率差值为 3.5%和 4.5%.

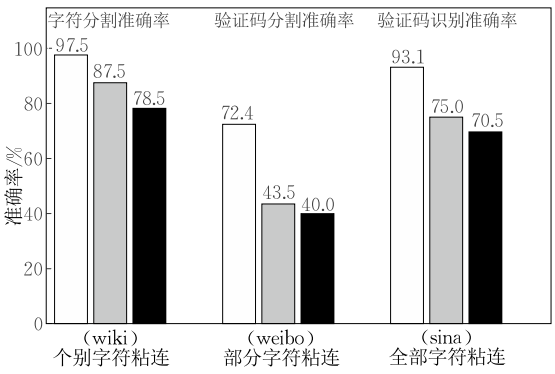


图 21　字符分割准确率、验证码分割准确率和验证码识别准确率对比图

表 5 为本文识别系统对于上文数据集 2 中所有验证码的识别准确率统计表.从表中可以看出京东验证码的识别率最高，而百度验证码的识别率最低.京东验证码较高的识别率是因为验证码中字符未出现扭曲和旋转变换，背景较为简单，可极大的发挥出 RBNet 和 CSNet 的效果.对于百度验证码识别率较低的原因是该验证码在设计时不仅对字符进行了扭曲，而且对字符进行了较大程度的粘连.较大程度的粘连限制了 CSNet 的拉伸效果，使得最终识别准确率为 25%.

表 5　数据集 2 验证码识别准确率展示表

验证码	本文识别准确率/%
京东	86.0
支付宝	36.5
易贝	29.0
微软	31.0
微博	40.0
百度	25.0
360	64.5
维基百科	78.5
搜狐	70.5
新浪	70.5

表6呈现的是本文系统识别准确率和研究[8,14,36-37]识别准确率对比表.其中文献[37]为整体识别方案，本文分割识别方案对于除 Yahoo(2016)之外的验证码都有更好的识别准确率.文献[14]中的验证码多为含有复杂背景干扰但未对字符进行粘连和扭曲变换，本文预处阶段 RBNet 可以有效的清除背景干扰信息，平均破解率达 86.05%.本文方法对文献[8]中的 QQ 和维基百科的破解率都有明显的提升.其

表 6　验证码识别准确率对比表

验证码	准确率/%		证码	准确率/%		证码	准确率/%		验证码	准确率/%	
	CCS (2011)[14]	本文		NDSS (2016)[8]	本文		USENIX Security (2014)[36]	本文		Science (2017)[37]	本文
Megaupload	93	97.5	Baidu(2016)	46.6	96.5	reCaptcha(2013)	22.3	45.0	PayPal	57.1	72.5
Blizzard	70	83.5	QQ	56.0	99.0	Baidu(2013)	55.2	99.0	reCaptcha(2011)	66.6	77.5
Authorize	66	88.0	Taobao	23.4	52.0	reCaptcha(2011)	22.7	77.5	Yahoo(2016)	57.4	45.0
Captcha.net	73	95.5	Sina	9.4	75.0	Baidu(2011)	38.7	88.0			
NIH	72	96.0	reCaptcha(2011)	77.2	77.5	Wikipedia	28.3	87.5			
Reddit	42	90.0	Amazon	25.8	96.0	CNN	51.1	95.5			
Digg	20	84.5	Wikipedia	23.8	87.5						
Wikipedia	25	87.5	Microsoft	16.2	36.0						

中 QQ 验证码准确率提升至 99%，主要是因为文中 QQ 验证码为空心字体，RBNet 可以高效的对字符进行填充，再通过 CSNet 对字符间距进行拉伸，从而很大程度上优化了分割过程. 维基百科验证码识别准确率有较大的提升，说明 CSNet 对于字符间距的有效拉伸对识别准确率有显著的提高. 文献[36]中百度(2013)识别率较高的原因也是因为 RBNet 对于空心字符有效填充. 相比于文献[37]中 Yahoo(2016)验证码有所减低，其原因是 Yahoo(2016)验证码中字符较多且粘连程度较大，影响了 CSNet 的拉伸效果. 因部分网站验证码系统进行了升级，对于真实验证码无法获取的网站，本文采用程序模拟验证码进行测试.

对于含有背景干扰信息且字符粘连的验证码诸如新浪和 360 的验证码，本文预处理阶段分别使用了 RBNet 和 CSNet 两个网络依次进行去干扰和字符拉伸操作. 表 7 呈现了若将两个网络合并和分开处理准确率的变化. 实验表明对于含有背景干扰且字符粘连的验证码使用两个网络分别进行操作的效果要优于使用同一个网络的效果.

表 7 网络使用探究验证码识别准确率对比表

验证码	合并网络/%	RBNet+CSNet/%
新浪	65.0	70.5
360	63.5	64.5

5 验证码保护策略

在上文识别实验中详细分析了多种验证码特点，并对其安全性进行了评估. 本节通过详细梳理上文的实验过程及实验结果，提出五点有效的验证码保护策略.

5.1 验证码安全特征分析

本文将验证码设计时所采用的安全特征分为有效的安全特征和无效的安全特征.

在设计验证码时有效的安全特征如下所述：

(1) 字符之间相互粘连

字符之间相互粘连是设计时使用最多的安全特征，Alexa 发布的 TOP 50 的网站使用了字符粘连策略的验证码占文本验证码的 90%以上. 文献[21]中也提到字符粘连是防止字符分割最有效的方式，字符之间粘连程度和粘连字符的个数都对验证码的安全性有着较大的影响.

(2) 较多的字符数量

Yan 等人的研究[13]首次提出验证码中字符数量对验证码的安全性较大的影响，Bursztein 等人[14]随后的研究中也说明了这一点. 对于本文中的预处理阶段，较多的字符会影响 CSNet 的拉伸效果. 较多的字符也可以增加字符分割难度，影响分割的准确率. 在识别阶段，整体识别准确率由单个字符的识别结果组合而成，随着字符个数的增加，整体识别准确率也会受到影响.

(3) 采用特殊的字体

使用机器学习的方法对验证码进行识别，需要大量的验证码和对应的验证码内容进行监督训练. 现在多数网站都无法直接获得大量真实的验证码，即便是获得了大量真实的验证码也没有对应的训练标签. 若采用模拟验证码进行训练，则前提是可以获得和真实验证码相似的字体，若没有相似的字体，即便生成的验证码中字符的大小，分布都与真实验证码非常相似，识别准确率也会很难保证.

在设计验证码时无效的安全特征如下所述：

(1) 相似的背景图案

在本文的预处理阶段，当使用的背景为一种或者几种相似的图案时，使用 RBNet 便可以干净的去除(图 22 为京东和 360 验证码预处理效果图). 因此，验证码使用一种或几种相似的背景图案作为干扰信息不会提高验证码的安全性.

图 22 验证码预处理效果图

(2) 空心字符

由于空心字符轮廓线较细的特点，部分验证码在设计时使用空心字符来代替实心字符. 当空心字符验证码含有干扰线时，较细的字符轮廓线会导致形态学腐蚀等方法无法使用，所以空心字符中的干扰信息更难去除. 在分割阶段，对空心字符进行垂直投影，很难通过像素直方图来找到字符连接点，增加字符分割的难度. 即使有效分割之后，空心字符的特征点较少，识别的准确率也低于实心字符. 为降低空心字符的干扰，研究者使用图像学的方法把空心字符填充成实心字符进行处理. 然而对于含有干扰线的空心字符验证码，现存的填充算法有效性仍有待提高. 本文采用条件生成式对抗网络的思路在去除

干扰信息的同时对字符进行填充，实现效果如图 22 中 360 验证码所示. 通过本文方法可清晰的将空心字符验证码转换成实心字符验证码，使用空心字符的优势消失. 因此，使用空心字符不会提高验证码的安全性.

5.2　验证码保护策略

在对识别算法进行大量的研究之后，为进一步提高验证码的安全性，除了使用上文中提到有效的安全策略之外，本文还给出了如下几点设计建议.

(1) 混合使用多种特殊字体.

使用多种字体是用来增加识别难度，使用特殊的字体则是防止攻击者生成与真实验证码相似的程序模拟验证码. 攻击者无法生成相似的样本进行训练，识别准确率就会大大降低.

(2) 重叠字符笔画

重叠字符笔画是对字符粘连的一种优化，若字符笔画重合则无法进行分割，即使分割了有一部分也会是残缺字符，识别准确率必然降低.

(3) 字符中添加断裂线

在验证码中添加断裂线(如图 23 所示)，是用来防止基于投影算法的分割算法. 经过这样处理之后，像素直方图上会出现很多值为零的点，这样就可以混淆分割点的选择，从而提高分割难度.

图 23　添加断裂线效果图

(4) 混合使用多种不同风格的验证码

采用机器学习的方法识别验证码，对于不同风格的验证码需要进行多次训练. 若验证码系统中含有多种不同风格的验证码，在破解时则需要进行多次训练. 这样不仅提高了系统的复杂性，还增加了破解需要的时间.

(5) 对抗样本验证码生成策略

对抗样本是一个通过对图片中像素进行细微调整，使得人眼难以察觉其变化且神经网络识别出错的输入样本. 图 24 则是通过 FGSM[38] 算法在 Amazon 的验证码中添加了干扰信息，从而使得之前训练好的模型识别出错. 通过不定期向验证码生成系统中添加对抗样本可大大降低验证码的破解率.

图 24　Amazon 对抗样本效果图

6　结论与展望

本文提出了一种基于条件生成式对抗网络的文本类验证码识别方法，该方法在预处理阶段两次运用条件生成式对抗网络，将含有干扰信息且字符扭曲、旋转和粘连的验证码转换为“背景更干净，字符间距更大”的验证码，为字符分割提供了有利的条件. 对于分割识别方案，通过本文预处理方法，可以很大程度上提高字符分割的准确率，从而提高验证码识别准确率. 本文大量实验表明：本文方法对于含有干扰信息且字符复杂变换的验证码有较高的识别准确率和较短的识别时间.

未来的工作有：提高 CSNet 对于字符重叠较多验证码的拉伸效果，减少验证码分割准确率和验证码识别准确率之间的损失. 通过以上两点研究，将会整体提高本文方法的识别效率并扩大其适用范围.

参　考　文　献

[1] Lorenzi D, Uzun E, Vaidya J, et al. Towards designing robust CAPTCHAs. Journal of Computer Security, 2017, 26(2): 1-30

[2] Zhu B B, Yan J, Bao G, et al. Captcha as graphical passwords—A new security primitive based on hard AI problems. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 891-904

[3] Aiken W, Kim H. POSTER: DeepCRACk: Using deep learning to automatically crack audio captchas//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. Incheon, Korea, 2018: 797-799

[4] Sivakorn S, Polakis I, Keromytis A D. I am robot: (deep) learning to break semantic image captchas//Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Saarbrücken, Germany, 2016: 388-403

[5] Shah M, Harras K. Hitting three birds with one system: A voice-based CAPTCHA for the modern user//Proceedings of the 2018 IEEE International Conference on Web Services (ICWS). San Francisco, USA, 2018: 257-264

[6] Choi J, Oh T, Aiken W, et al. POSTER: I can't hear this because i am human: A novel design of audio CAPTCHA system//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. Incheon, Korea, 2018: 833-835

[7] Bursztein E, Beauxis R, Paskov H, et al. The failure of noise-based non-continuous audio captchas//Proceedings of the 2011 IEEE Symposium on Security and Privacy. Nice, France, 2011: 19-31

[8] Gao H, Yan J, Cao F, et al. A simple generic attack on text captchas//Proceedings of the Network and Distributed System Security Symposium. San Diego, USA, 2016: 220-232

[9] Ye G, Tang Z, Fang D, et al. Yet another text captcha solver: A generative adversarial network based approach//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada, 2018: 332-348

[10] Yin Long, Yin Dong, Zhang Rong, et al. A recognition method for distorted and merged text-based CAPTCHA. Pattern Recognition and Artificial Intelligence, 2014, 27(3): 235-241(in Chinese)
(尹龙, 尹东, 张荣等. 一种扭曲粘连字符验证码识别方法. 模式识别与人工智能, 2014, 27(3): 235-241)

[11] Chellapilla K, Simard P Y. Using machine learning to break visual human interaction proofs (HIPs)//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 265-272

[12] Yan J, El Ahmad A S. Breaking visual captchas with naive pattern recognition algorithms//Proceedings of the 23rd Annual Computer Security Applications Conference (ACSAC 2007). Miami Beach, USA, 2007: 279-291

[13] Yan J, El Ahmad A S. A low-cost attack on a Microsoft captchas//Proceedings of the 15th ACM Conference on Computer and Communications Security. Alexandria, USA, 2008: 543-554

[14] Bursztein E, Martin M, Mitchell J. Text-based CAPTCHA strengths and weaknesses//Proceedings of the 18th ACM Conference on Computer and Communications Security. Chicago, USA, 2011: 125-138

[15] Gao H, Wang W, Qi J, et al. The robustness of hollow CAPTCHAs//Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. Berlin, Germany, 2013: 1075-1086

[16] Li Xing-Guo, Gao Wei. Segmentation method for merged characters in CAPTCHA based on drop fall algorithm. Computer Engineering and Applications, 2014, 50(1): 163-166(in Chinese)
(李兴国, 高炜. 基于滴水算法的验证码中粘连字符分割方法. 计算机工程与应用, 2014, 50(1): 163-166)

[17] Congedo G, Dimauro G, Impedovo S, et al. Segmentation of numeric strings//Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Canada, 1995: 1038-1041

[18] Lv Ji. Study on the Recognition of Validation Code Based on Neural Network [M. S. dissertation]. Huaqiao University, Xiamen, Fujian, 2015(in Chinese)
(吕霁. 基于神经网络的验证码识别技术研究[硕士学位论文]. 华侨大学, 厦门, 福建, 2015)

[19] Liu Huan, Shao Wei-Yuan, Guo Yue-Fei. Research on captcha recognition with convolutional neural networks. Computer Engineering and Applications, 2016, 52(18): 1-7 (in Chinese)
(刘欢, 邵蔚元, 郭跃飞. 卷积神经网络在验证码识别上的应用与研究. 计算机工程与应用, 2016, 52(18): 1-7)

[20] Wang Lu. The Research on Captcha Recognition Technology [M. S. dissertation]. University of Science and Technology of China, Hefei, 2011(in Chinese)
(王璐. 验证码识别技术研究[硕士学位论文]. 中国科学技术大学, 合肥, 2011)

[21] Huang S Y, Lee Y K, Bell G, et al. A projection-based segmentation algorithm for breaking MSN and YAHOO CAPTCHAs//Proceedings of the International Conference of Signal and Image Engineering. Heidelberg, Germany, 2008: 1135-1149

[22] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[23] Wang K F, Gou C, Duan Y J, et al. Generative adversarial networks: The state of the art and beyond. Acta Automatica Sinica, 2017, 43(3): 321-332

[24] Lecun Y L, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[25] Xu Ke. Study of Convolutional Neural Network Applied on Image Recognition [M. S. dissertation]. Zhejiang University, Hangzhou, 2012(in Chinese)
(许可. 卷积神经网络在图像识别上的应用的研究[硕士学位论文]. 浙江大学, 杭州, 2012)

[26] Colomer A, Naranjo V, Angulo J. Colour normalization of fundus images based on geometric transformations applied to their chromatic histogram//Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China, 2017: 3135-3139

[27] Zhang Jing-Lei, Zhang Yun-Fei. Behavior understanding of moving objects based on improved watershed algorithm. Computer Engineering and Design, 2015, 36(7): 1840-1844 (in Chinese)
(张惊雷, 张云飞. 基于改进分水岭算法的运动目标行为理解. 计算机工程与设计, 2015, 36(7): 1840-1844)

[28] Liu Wan-Jun, Zhao Yong-Gang, Min Liang. Automatic image segmentation method based on k-means and FCM. Computer Engineering and Applications, 2015, 51(16): 199-203(in Chinese)
(刘万军, 赵永刚, 闵亮. 结合 k-means 的自动 FCM 图像分割方法. 计算机工程与应用, 2015, 51(16): 199-203)

[29] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Swiss, 2015: 234-241

[30] Li C, Wand M. Precomputed real-time texture synthesis with Markovian generative adversarial networks//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 702-716

[31] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2536-2544

[32] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1125-1134

[33] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105

[34] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[35] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 4278-4284

[36] Bursztein E, Aigrain J, Moscicki A, et al. The end is nigh: Generic solving of text-based captchas//Proceedings of the USENIX Conference on Offensive Technologies. San Diego, USA, 2014: 3

[37] George D, Lehrach W, Kansky K, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. Science, 2017, 358(6368): eaag2612

[38] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-11

TANG Zhan-Yong, Ph. D., professor. His main research interests include software security and protection, wireless sensor network security.

TIAN Chao-Xiong, M. S. His main research interest is information security.

YE Gui-Xin, Ph. D., lecturer. His main research interests include privacy and software security, authentication.

LI Jing, M. S., lecturer. Her main research interest is software security.

WANG Wei, Ph. D., associate professor. Her main research interest is mobile computing and network information security.

GONG Xiao-Qing, Ph. D., associate professor. Her main research interests include software security and workflow.

CHEN Xiao-Jiang. Ph. D., professor. His main research interest is wireless sensor network, software security and protection.

FANG Ding-Yi, Ph. D., professor. His research interests include network and information security, software security and protection, key technologies of wireless sensor network.

Background

This paper is related to authentication. Text-based captchas are widely used to distinguish humans from automated computer programs. Despite numerous alternative schemes to text-based captchas have been proposed, current many popular websites still use text-based captchas to protect against bots. Due to the wide usage of text-based captchas, a compromise on them could result in serious consequences.

So far, researchers have proposed many methods for automatically recognizing text-based captchas. However, on one hand, many of previous approaches are focus their attention on a few specific captcha schemes, tuning the their attacking models to other schemes requires heavy expert involvement. On the other hand, some prior attacking methods requires millions of target text captchas to learn a deep learning models. But collecting and labelling so many captchas requires a labor-intensive and time-consuming process to construct. In the past decades, text-based captchas are keeping evolving and have become more robust, such as current captcha schemes have introduced more security features such as more complex confusing background as well as rotate, distorted and overlapping characters.

In this paper, we present a generic generative adversarial network based attack on text-based captchas. Unlike previous machine-learning-based attacks, our approach does not require

a large volume of real captchas to learn an effective solver, we significantly reduces the number of real captchas needed. This is achieved by first using CGAN to remove the background interference information and stretch the character spacing of the Captchas. Then we use the optimization of segmentation algorithms to segment the stretched Captchas effectively, and use GoogLeNet to perform single character identification. In addition, it is difficult to obtain a large number of real Captchas at a low cost. This paper designs a program to simulate these real Captchas for network training, and the training cost is far lower than other existing methods and the training effects are the same. We evaluate our approach by applying it to a total of 10 text captcha schemes which are currently being used by most of top-50 popular websites ranked by alexa.com as of March, 2018. The experimental results demonstrate that our attacking method can significantly outperform all previous state-of-the arts.

We hope that the results of our work can encourage the security community to revisit the design and practical usage of text-based captchas.

This work is supported by the National Natural Science Foundation of China (61672427, 61972314), the Shaanxi International Cooperation Program (2017KW-008), the Shaanxi International Cooperation Program (2019KW-009), the Shaanxi Provincial Key Research and Development Program (2017GY-191), and the Shaanxi Province Innovation Funded by the Team (2018SD0011).