

基于 PSO-CNN 的验证码识别算法研究

李建平,王 钊

(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:伴随着互联网的高速发展,非法用户恶意攻击网站、恶意注册、暴力破解用户密码等事件也随之而来。为了解决这些网络安全问题,作为网络安全第一道防线的验证码技术应运而生。但在实现自动登录合同管理系统的过程中,验证码自动化识别一直是个技术难点,验证码自动化识别准确率直接影响了业务处理效率,故此提出了一种基于 PSO-CNN 的验证码识别方案。针对一万张验证码图片的数据集进行灰度化、二值化以及降噪三步数据预处理之后,通过 PSO 优化算法在卷积神经网络训练数据集的过程中找出最佳的网络层数和卷积核大小。经过反复的实验,结果表明基于 PSO-CNN 的验证码识别算法对数字与字母混合验证码识别准确率可达 96.26%,为合同管理系统实现自动登录提供了可靠的技术支持。

关键词:粒子群优化算法;卷积神经网络;验证码;数据预处理;Tesseract

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2022)09-0051-05

doi:10.3969/j.issn.1673-629X.2022.09.008

Research on Verification Code Recognition Algorithm Based on PSO-CNN

LI Jian-ping, WANG Zhao

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: With the rapid development of Internet, illegal users malicious attacks on websites, malicious registration, violent cracking of user passwords and other events have followed. In order to solve these network security problems, verification code technology, as the first line of defense of network security, came into being. However, in the process of realizing automatic login to the contract management system, automatic identification of verification code has always been a technical difficulty. The automatic identification accuracy of verification code directly affects the business processing efficiency. Therefore, a verification code recognition scheme based on PSO-CNN is proposed. After the three-step data preprocessing of grayscale, binarization and noise reduction for 10000 verification code images, the best number of network layers and convolution kernel size are found in the process of convolutional neural network training data set through PSO optimization algorithm. The results of repeated experiments show that the recognition accuracy of mixed digital and letter verification code based on PSO-CNN has reached 96.26%, which provides reliable technical support for automatic login of system.

Key words: particle swarm optimization (PSO); convolutional neural network (CNN); verification code; data preprocessing; Tesseract

0 引言

验证码是分辨用户是机器还是人的一种公共全自动算法程序。作为互联网领域保证网络安全的常用手段,验证码技术的产生极大程度上阻止了非法用户暴力破解用户密码、频繁登录与注册等恶意行为,有效地保证了网站内部数据的安全,避免了大量僵尸用户的产生^[1]。但是对于在机器人流程自动化领域,为实现自动化处理业务就必须实现验证码自动化识别。然而数量巨大且抽象的验证码,人工识别都很难达到高的

准确率,更何况使用机器识别。这就间接导致了实现自动化处理业务似乎变成了一个无法实现的难题。

传统的验证码识别方法一般基于 OCR 技术。OCR 即光学字符识别,其实现原理是利用电子设备检查图形的亮暗的形状来翻译成计算机字符的过程。OCR 识别验证码之前需要对源数据进行预处理。例如二值化、行切分、连通域分析等。虽然对于比较规整的字符序列采用 OCR 技术也会取得较好的识别效果,胡晓辉^[2]利用 Tesseract 光学字符识别引擎实现了对

收稿日期:2021-10-10

修回日期:2022-02-16

基金项目:国家自然科学基金重点项目(61933007)

作者简介:李建平(1976-),男,博士,教授,CCF 会员(38468M),研究方向为智能计算;通讯作者:王 钊(1994-),女,硕士,研究方向为智能计算。

验证码的识别,并取得了较好的识别效果。但是随着大量具有抽象、粘连、噪点等特征图像验证码的出现,OCR 技术识别效果却收效甚微^[3]。其缺点表现如下:(1)对于复杂验证码图像二值化处理会造成信息缺失;(2)人工干预过多。对方法参数的设定完全依赖于当事人实践经验;(3)OCR 识别灵活性差,导致对复杂验证码图像处理修改空间急速变小;(4)大量的电子识别设备需要高额成本;(5)英文字母识别率低下。

直至本世纪初卷积神经网络理论的逐步成熟使得验证码自动化识别有了新的突破。相比于其他的验证码识别技术,卷积神经网络拥有更加强大的容错能力、自学习能力、泛化能力和并行处理能力^[4]。改变全连接层神经元个数所构建的基于深度卷积神经网络 AlexNet 多任务验证码识别模型^[5]以及通过 Inception 模块替换 Google-net 的卷积层所构建的基于端到端的验证码识别模型^[6]等都取得了不错的识别效果。故针对传统验证码识别技术上的不足,该文提出了一种基于 PSO-CNN 的验证码识别技术。验证码识别模型的建立通过四个部分来完成,分别为验证码源数据预处理、搭建卷积神经网络识别模型、寻优识别模型参数、评价验证码识别效果。

1 相关理论

针对合同管理系统的验证码识别,采用粒子群优化算法与卷积神经网络相结合的方式。即利用卷积神经网络作为训练识别模型的基础,而粒子群寻优算法则是为了找出最优的识别模型参数。故为获得优化后的验证码识别模型,该文以卷积核大小和网络层数作为粒子群寻优的问题域,将验证码识别的准确率作为寻优算法的标准,并最终返回某一迭代范围内的最优的卷积核大小和网络层数大小参数。进而得到优化后的验证码识别模型。

1.1 卷积神经网络(CNN)

卷积神经网络作为一种前馈神经网络是进行大型图像处理的利器,其网络层级结构由数据输入层(原始图像预处理)、卷积层(卷积神经网络中重要的一层,完成提取特征的过程)、ReLU 激励层(将输出结果做非线性映射。ReLU 函数具有收敛快、计算梯度快的特点)、池化层(用于压缩数据和参数的量,减少过拟合现象的发生)和全连接层(“分类器”的作用)依次构成^[7]。

1.2 粒子群算法(PSO)

基于上述卷积神经网络虽然为验证码图像识别提供了新的思路,但是在卷积神经网络训练参数的设置完全取决于技术人员的实际经验,这就直接导致不同的参数将会得到不同的验证码识别准确率。甚至有时

因为参数设置的原因,不同的验证码识别模型预测准确率相差甚大。故从卷积核与卷积神经网络层数这两个方向入手,该文利用粒子群优化算法在卷积神经网络构建验证码预测模型的过程中寻找出适合于当前验证码识别模型的最优参数。

为解决实际工程中的最优化问题,智能算法应运而生。遗传算法(GA)、退火算法(SA)、粒子群优化算法(PSO)等都是智能算法中的代表算法。其中粒子群优化算法因简单易行、收敛速度快、设置参数少的优势备受学界的青睐^[8]。该算法思想是受鸟类或鱼类觅食的行为的启发。该文将寻优的参数视为粒子群优化算法中的“粒子”,将这些“粒子”置于解空间中进行迭代搜索,在搜索的过程中利用“粒子”之间的协作与信息共享来最终得到验证码识别模型的参数最优解。

在粒子群优化算法中,每一个寻优粒子赋予记忆功能。其具有两个属性,分别为速度与方向。速度代表的是寻优粒子在解空间中活动的快慢。寻优粒子后续状态由自身的飞行经验和同伴的飞行经验共同决定。寻优粒子的速度不宜过快,速度过快将很可能导致错过全局最优解。方向代表寻优粒子在解空间中活动的方向^[9]。每一个粒子所处位置与最优解或满意解之间的偏差通过测度函数来进行衡量。其中寻优粒子的速度更新公式为:

$$v_i^m = wv_i^{(m-1)} + c_1 * \text{rand}() * (\text{pbest}_i - x_i^{(m-1)}) + c_2 * \text{rand}() * (\text{gbest}_i - x_i^{(m-1)}) \quad (1)$$

通过上面的公式可以看出,寻优粒子的速度取决于三方面:(1)粒子先前速度;(2)粒子*i*当前位置与本身历史最好位置之间的距离;(3)粒子当前位置与种群最好位置之间的距离。其中 v_i^m 表示第*m*次迭代粒子在解空间中的飞行速度。 $\text{rand}()$ 表示介于(0,1)之间的随机函数。 pbest_i 表示寻优粒子经过的最好位置, gbest_i 表示种群经历的最好位置。 c_1 、 c_2 表示加速度常数。 w 即非负惯性参数,决定了全局寻优能力的强弱。

粒子*i*的位置更新公式为:

$$x_i^m = x_i^{(m-1)} + v_i^{(m-1)} \quad (2)$$

其中, x_i^m 表示第*m*次迭代粒子*i*的当前位置。

故验证码识别模型的寻优参数步骤如下:

(1)设定验证码识别模型寻优种群规模,随机初始化每个寻优粒子以及最大迭代次数。其中种群规模的设定不宜过小,太小则陷入局部最优解的可能性就会加大。粒子群优化算法优化能力强,当种群数数目增至一定数量时,再增长所寻优结果将不会发生太大变化。

(2)根据测度函数计算每一个寻优粒子的适应度。其中将验证码识别模型的准确率作为测度函数的返回值^[10]。

(3)更新寻优粒子本身历史最好位置、种群历史最好位置、粒子的速度与位置。

(4)当达到最大迭代次数或者满足设定最小误差时,停止迭代即完成验证码识别模型的建立。否之返回步骤2继续寻找最优解。

2 基于 PSO-CNN 验证码识别模型构建

验证码预测模型构建的过程大致分为三个部分:(1)向前传播;(2)损失函数值计算;(3)反向传播。识别模型不断调整权重和偏置的值,当误差小于某一个设定误差值时,则停止迭代完成验证码预测模型的构建。验证码预测模型的构建流程如图1所示。

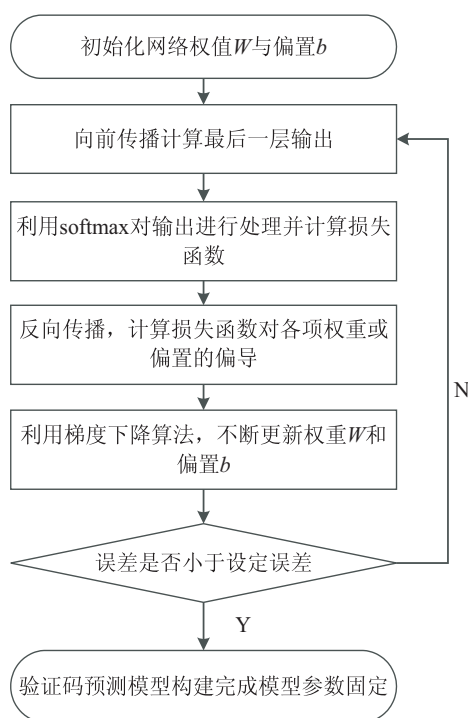


图1 验证码预测模型的构建流程

(1)向前传播。向前传播分为:卷积层向前传播、池化层向前传播、全连接层向前传播。卷积层向前传播指的是通过事先设定的卷积核大小对输入验证码图像进行卷积处理。传播计算公式为:

$$z_j^l = \sum_{i=1}^M a_j^{(l-1)} * w_{ij}^l + b_j^l, a_j^l = \sigma(z_j^l) \quad (3)$$

其中, M 为输入节点 i 的总数, w_{ij} 为总的卷积核个数。 z_j^l 表示总输入通道数为 M 的节点 a_i , 经过第 l 层网络层处理后产生第 j 个输出节点, 再经过激活函数, 成为下一网络层的第 j 个输入通道的特征节点 a_j^l 。

池化层向前传播。池化层也叫下采样层, 该层将卷积层提取的特征数据利用最大池化算法或者平均值池化算法^[11], 降低特征数据的计算难度, 有效防止了过拟合现象的发生。

$$z_j^l = \text{down}(a_j^{(l-1)}), a_j^l = \sigma(z_j^l) \quad (4)$$

其中, z_j^l 表示节点 a_j 经过第 l 层池化操作后的输出结果。

全连接层向前传播。经过卷积和池化操作之后将特征数据输入至全连接层进行输入数据的分类, 并建立预测模型。

(2)损失计算。损失函数用于衡量计算实际分类结果与标签结果之间的偏差。偏差越大说明模型分类效果越差, 反之偏差越小分类结果相对越好。损失函数公式为:

$$E^p = \sum_{p=0}^P \frac{1}{2} \|a^L - y\|^2 \quad (5)$$

其中, E^p 表示预测模型损失偏差的和, a^L 表示第 L 层的输出节点, y 表示标签结果, P 表示有数据的组数。

通过对预测模型损失函数的计算, 再利用梯度下降法不断更新权重 W 和偏置 b 。参数更新公式为:

$$(W, b) = \text{argmin}(E^p(W, b)) \quad (6)$$

$$(W, b)^{(n+1)} = (W, b)^n - \alpha \cdot \text{grad}(E^p(W, b)^n) \quad (7)$$

其中, α 表示学习率, 学习率表示模型对最优参数 W , b 寻找的仔细程度。grad 表示梯度下降函数。

(3)反向传播算法。当预测值与标签值误差值大于设定误差时, 将误差值一层一层返回, 计算出每一层的误差并更新权重参数。

3 数据预处理方法

获取吉林某油田合同管理系统登录页面的验证码数据集。源数据集包含三个特征:(1)抽象。字母、数字并不是规整排列的, 而是歪歪扭扭并行排列;(2)粘连。有相当一部分的验证码相邻的字母或数字之间相互连接在一起;(3)含噪点。每一张验证码图片包含了干扰识别准确率大的噪点。其原始图片效果如图2所示。



图2 未处理的验证码

对于原始数据集的处理, 该文主要分为灰度化处理、二值化、降噪三步。灰度化指的是将彩色图片转化为灰度图片的过程。由于彩色图像具有 R 、 G 、 B 三个通道, 直接用彩色图片进行训练将会导致建模时间过长, 既耗时又耗力。故将源数据集进行灰度化操作转化为单色道的灰度图片。灰度化操作一般包含三种算法: 最大值法、平均值法和加权平均值法。加权平均值法^[12]灵活度高、自定义性好, 故使用该方法完成对图像的灰度化操作。其公式如下:

$$R = G = B = \frac{\omega_1 R + \omega_2 G + \omega_3 B}{3} \quad (8)$$

其中, ω_1 、 ω_2 和 ω_3 分别为 R 、 G 、 B 的权值。取不同的权重参数可以得到不同的灰度图像。 ω_1 、 ω_2 、 ω_3 分别取值 0.299、0.587、0.114。其得到的灰度验证码图片如图 3 所示。



图 3 灰度化后的验证码

为了更好地提取图像中的信息,增加验证码识别的准确率,需要对图片进行二值化处理。二值化^[13]处理顾名思义指的是按照某一个灰度阈值将灰度化后的验证码图像像素点划分为白色(0)或者黑色(255)两部分。故经过二值化后的图片只会呈现黑白两种效果。二值化处理的关键点在于灰度阈值的选择。该文选用全局阈值的方法将灰度阈值设定为 200。其二值化处理后的效果如图 4 所示。



图 4 二值化后的验证码

降噪操作进一步降低了识别模型的数据计算难度,剔除了大部分二值化验证码图像中的干扰点。即降噪^[14]指的是删除图像中的干扰点,保留图像主体特征信息的过程。该文选用非局部均值降噪^[15]的方法对二值化图像进行降噪处理。其处理结果如图 5 所示。



图 5 降噪后的验证码

4 实验结果及分析

4.1 实验软硬件环境

硬件环境:Windows 7 旗舰版。

软件环境:PyCharm 2018; Jupyter Notebook; Tesseract 5.0.0; Python 3; Pytorch。

4.2 数据集及预处理

数据样本来源:选取吉林某油田合同管理系统登录界面的验证码图片作为数据集。

数据样本类型及构成:数字与字母组合。

样本数量及划分:10 000 张,按 8 : 2 分为训练集和测试集。

实验数据、实验的软硬件环境准备就绪后,针对验证码数据集的识别,将对比光学字符识别引擎 Tesseract 与粒子群优化-卷积神经网络识别这两个方法所建立的验证码识别模型。

首先 Tesseract 是由 HP 实验室开发的一款 OCR 引擎。在不同的实际生产环境,定制不同引擎模板,通

过对数据的不断训练来获取验证码图像的识别模型^[2]。该实验利用其封装框架 PyTesseract 完成数据的预处理操作以及识别模型的建立。表 1 展示的是基于 Tesseract 识别合同管理登录系统验证码的准确率,表 2 展示的是数据集处理后不同类别的验证码识别效果。

表 1 基于 Tesseract 的合同管理登录系统验证码识别效果

指标	数据集处理前	数据集处理后
验证码识别准确率/%	11.3	60.5

表 2 不同类别的验证码识别效果

指标	数据集处理前		数据集处理后	
	纯数字验证码	非纯数字验证码	纯数字验证码	非纯数字验证码
验证码识别准确率/%	13.5	10.9	68.6	59.1

表 1 的实验结果表明,基于 Tesseract 的合同管理登录系统验证码识别模型在数据预处理后识别准确率有着较大的提升。其数据处理前验证码识别准确率为 11.3%,经过灰度化、二值化、去噪点数据预处理之后识别准确率可达 60.5%。表 2 的实验表明,Tesseract 对合同管理登录系统的纯数字验证码比非纯数字验证码有着较好的识别效果。可以看出无论是数据处理前还是处理后,验证码识别准确率都难以达到令人满意的结果,而且在 Tesseract 识别的过程中识别出乱码字符、空字符的情况也时有发生。故经过以上的实验,可以得出结论:基于 Tesseract 实现合同管理登录系统验证码识别的方案是不可行的。

4.3 参数设置及评价指标

利用 PSO 算法与 CNN 相结合的方法,找出卷积神经网络识别合同管理系统验证码最优的卷积核和网络层数。在构建验证码预测模型的过程中,基本参数设置如表 3 所示。

表 3 PSO-CNN 验证码预测模型基本参数设置

参数类型	参数值
PSO 非负惯性参数 w	0.8
PSO 加速度常数 c_1 、 c_2	0.5
PSO 最大迭代次数 max_iter	15
CNN 卷积步长 stride	1
CNN 填充模式	Zero
CNN 学习速率	0.001

4.4 性能分析

经过 168.5 小时的模型训练后得到了优化后的验证码预测模型,其平均粒子群优化算法每迭代依次需

要花费 10 小时左右。由于计算机本身的限制,随着程序运行时间的增加其模型每次迭代所花费的时间也在逐步增加。基于 PSO-CNN 所构建的验证码识别模型的识别效果如表 4 所示。

表 4 基于 PSO-CNN 的验证码识别效果 %

Layers	Kernels			
	3 * 3	5 * 5	7 * 7	9 * 9
3	94.26	91.03	90.47	85.91
4	96.26	90.01	95.23	86.01
5	89.25	86.52	86.43	77.25

从表 4 的实验结果可以看出,不同的卷积神经网络参数所建立的验证码预测模型其识别准确率有着较大的差异。文中初始验证码识别模型参数设置为 5 * 5 卷积核以及四层网络结构,即在没有使用寻优算法的情况下单纯利用卷积神经网络识别验证码的准确率为 90.01%。经过寻优算法优化后所构建的验证码预测模型其验证码识别准确率最高可达 96.26%。就卷积核而言,随着卷积核的增大验证码识别准确率由于无法提取更多的特征信息,导致识别准确率总体趋势是下降的。从网络层数方面来说,随着网络层数的增加识别准确率有了明显的提高,但是当网络层数为 5 时,其所构建的验证码识别模型出现了不同程度的过拟合状态,导致验证码识别准确率出现了不同程度的下降。基于以上分析,图 6 展示了 Tesseract、CNN、PSO-CNN 三种方式的识别准确率的对比结果。

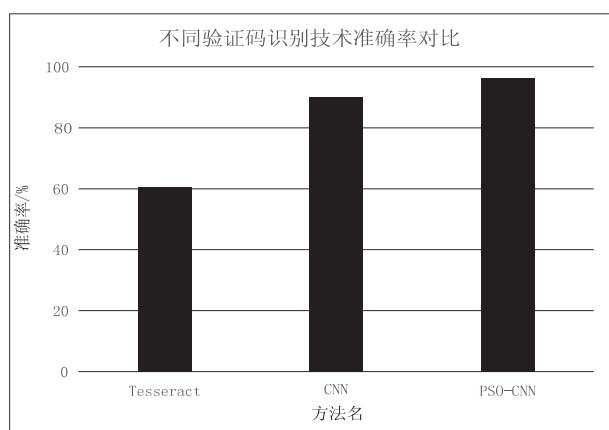


图 6 不同验证码识别技术准确率对比

5 结束语

针对传统的验证码识别技术识别粘连化、抽象化、含噪点的验证码识别效果差的情况,提出了一种 PSO-CNN 的验证码识别方案。该方案依托于卷积神经网络,在此基础之上利用粒子群优化算法找出解空间里面的最优模型参数。针对字母与数字混合的抽象复杂验证码,传统的 Tesseract 识别经常识别出乱码与空字

符,其识别最优准确率仅为 59.1%,通过文中方法所构建的验证码识别模型对于复杂验证码的识别最优准确率可达 96.26%,相对于传统的验证码识别方法其识别准确率提高了 37 个百分点,大大提高了复杂验证码识别的准确率。实验结果表明,该方案在验证码识别领域提供了可靠的技术支持,业务构建优化后的图像预测模型提供了一种崭新的思路。

参考文献:

- [1] 张锐,蔡艳林,陈夏裕,等.验证码的识别与改进[J]. 电脑编程技巧与维护,2021(5):117-119.
- [2] 胡晓辉. Tesseract 验证码识别探究[J]. 工业控制计算机, 2021,34(2):112.
- [3] 王日花. 基于深度学习的智能 OCR 识别关键技术及应用研究[J]. 邮电设计技术,2021(8):20-24.
- [4] MA J,ZHANG T,JING G,et al. Ground-based cloud image recognition system based on multi-CNN and feature screening and fusion[J]. IEEE Access,2020,8:1-4.
- [5] 于鹏. 基于深度卷积神经网络 AlexNet 的验证码识别研究[J]. 通讯世界,2018(1):66-67.
- [6] 崔新,白培瑞,张策,等. 一种基于端对端深度卷积神经网络的验证码识别方法[J]. 山东科技大学学报:自然科学版,2020,39(2):111-117.
- [7] ABIRAMI B, SUBASHINI T S, MAHAVAISHNAVI V. Gender and age prediction from real time facial images using CNN[J]. Materials Today:Proceedings,2020,33(7):4708-4712.
- [8] WU J,CHENG Y M,LIU C,et al. A BP neural network based on improved PSO for increasing current efficiency of copper electrowinning[J]. Journal of Electrical Engineering & Technology,2021,16(3):1297-1304.
- [9] 秦小林,罗刚,李文博,等. 集群智能算法综述[J]. 无人系统技术,2021,4(3):1-10.
- [10] KIM J,LEE K,CHOE J. Efficient and robust optimization for well patterns using a PSO algorithm with a CNN-based proxy model[J]. Journal of Petroleum Science and Engineering,2021,207:109088.
- [11] 刘万军,梁雪剑,曲海成. 不同池化模型的卷积神经网络学习性能研究[J]. 中国图象图形学报,2016,21(9):1178-1190.
- [12] 张国荣,刘炳君,付成丽. 基于 Python 和 CNN 的数字验证码识别[J]. 太原师范学院学报:自然科学版,2020,19(3):62-65.
- [13] 晋大鹏,张天心,刘涛. 基于 Python 和 CNN 的验证码识别[J]. 软件工程,2019,22(6):1-4.
- [14] 张娜娜,张媛媛,丁维奇. 经典图像去噪方法研究综述[J]. 化工自动化及仪表,2021,48(5):409-412.
- [15] 邢笑笑,王海龙,李健,等. 渐近非局部平均图像去噪算法[J]. 自动化学报,2020,46(9):1952-1960.