基于卷积神经网络的图片验证码识别研究

王 舜

（北京林业大学工学院北京100083）

摘要：【目的】在全国许多高校，报名全国大学生英语六级考试都是一个挑战。本文尝试以计算机手段实现报名网站的自动化登录以简化报名流程，关键难点在于识别登录界面的验证码图片，故开展本研究。【方法】首先收集验证码样本进行标注，制作数据集。然后使用敏捷开发的软件工程方法，以现有OCR模型为核心开发出软件原型，测试OCR的准确率并根据模型表现采取相应优化策略，软件原型开发成功后将识别方法替换为卷积神经网络，即用训练得到的CNN模型替换OCR方法，并使用单字符识别准确率和验证码图片识别准确率作为评估指标。软件主体为Python自动化测试工具selenium所实现的爬虫。【结果】使用程序从CET6报名网站爬取验证码图片并进行标注，获取有效样本725个，通过二值化验证码图片和过滤空格策略，OCR方法在该数据集的识别准确率提升至24.69%，具备了工程应用的可行性。以625张验证码图片作为训练集训练CNN在100张测试集图片上单字符识别准确率约48%，验证码识别准确率约5%，改进后的CNN准确率分别为50%和7%。通过数据增强将训练集扩增至50倍后，改进后的CNN单字符准确率可达94.5%，验证码识别准确率可达79%。【结论】分析认为1.验证码图片存在字符重叠的特点；2.字母不严格处于同一行对OCR干扰较大；3.OCR模型所识别字符集范围过大等原因使得OCR方法准确率较低。另一方面，改进后的CNN经数据增强的训练集训练得到的模型精度较高，单字符准确率达到94.5%，验证码识别准确率达到79%，远远超过OCR方法。经试验满足自动登录程序的要求。

关键词：卷积神经网络；验证码识别；深度学习

中图分类号：TP391.41       文献标志码：A       文章编号：xxxx-xxxx

**Abstract:** [**Objectives**] In many colleges and universities across our country, it is a challenge to register for CET6. This paper attempts to realize the automatic login of the registration website by computer to simplify the registration process. The key difficulty is to identify the verification code picture of the login page, so this research is carried out. [**Methods**] Firstly, collect the verification code samples for labeling, and make the data set. Then, using the software engineering method of agile development, the software prototype is developed with the existing OCR model as the core, the accuracy of OCR is tested, and the corresponding optimization strategy is adopted according to the model performance. After the software prototype is successfully developed, the recognition method is replaced by convolutional neural network, that is, the trained CNN model is used to replace the OCR model, Single character recognition accuracy and verification code picture recognition accuracy are used as evaluation indicators. The main body of the software is a crawler implemented by selenium, a python automated testing tool. [**Results**] using the program, 725 valid samples were obtained by crawling and then the pictures were labelled from the CET6 registration website. Through the binarization and space filtering strategy, the recognition accuracy of OCR method in the data set was improved to 24.69%, which has the feasibility of engineering application. Taking 625 verification code pictures as the training set, CNN is trained. On 100 test set pictures, the single character recognition accuracy is about 48%, the picture recognition accuracy is about 5%, and the improved CNN accuracy is 50% and 7% respectively. After the training set is expanded to 50 times through data enhancement, the improved CNN single character accuracy can reach 94% and the picture recognition accuracy can reach 79%. [**Conclusion**] The analysis shows that 1. The verification code picture has the characteristics of character overlap; 2. Characters are not strictly in the same line, which greatly interferes with OCR; 3. The accuracy is low due to the large range of character set recognized by OCR model. On the other hand, the improved CNN model trained by the data enhanced training set has high accuracy, with single character accuracy of 94% and picture recognition accuracy of 79%, which is far higher than that of OCR method. After test, it meets the requirements of automatic login program.

**Keywords:** convolutional neural network; verification code identification; deep learning

对全国许多在校大学生与高校本身来说，全国大学生英语四、六级考试（College English Test Band4 and Band6, CET4/CET6）报名[1]都是一个挑战。高校面临挑战的根本原因在于人均资源不足，所能提供的考位有限。因此许多高校采取特定措施缓解考位数量的压力，如限制报名名额[2]；或以一年级新生通过率不高为由，延迟一年级新生的报名资格，从而为二三四年级提供充足的考位[3]。因此对报名人数有限的六级考生而言，用“抢”这个动词描述报名过程是非常形象的。成功报名六级考试往往需要精准的计时、合适的设备与良好的上网环境和恰到好处的手速等缺一不可的“天时、地利、人和”三要素。这一过程既受外部因素的干扰如网络波动，又受制于报名者的自身条件如操作手速。因此，六级报名绝非一个简单的过程。本文尝试使用计算机技术建立一个报名网站自动化登录程序，以更好地了解CET6报名过程，并期望进一步形成CET6报名的指导方法。通过自动化测试工具selenium可以方便地模拟网页的点击、键盘输入等事件，因此研究的主要难点在于如何自动识别登录页面用于图灵测试的字符验证码。识别图片中的验证码本质上属于图像分类中的字符识别。基于传统机器学习的图像自动分类方法包括K邻近（K-nearest neighbor，KNN）[4]、贝叶斯网络[5]、支持向量机（support vector machine, SVM）[6]等，这类方法需要人为提取图像特征，选择合适的特征较为困难，识别效果有限。基于深度学习的图像自动分类方法属于机器学习的一个分支，通常采用卷积神经网络（convolutional neural network, CNN）实现。在本研究中采用敏捷开发方法，先选择现有光学字符识别方法(Optical Character Recognition, OCR)快速实现程序原型，再用CNN进行版本迭代。

本文的CNN模型在GitHub开源(<https://github.com/kafmws/deeplearning_demo>)，仓库中包括原始验证码数据集，数据集预处理工具、CNN实现等。并在checkpoints/目录下提供两个测试集准确率为75%和79%的模型文件。出于工程伦理角度考虑爬虫程序予以保留。

1. 数据与方法
   1. 数据集获取



图1 数据集制作

Fig 1 Captcha pictures with labels

如图1所示，使用程序从CET6报名网站登录页面批量保存验证码图片，统一为jpg格式，长宽分别为80和25像素。去除重复、无效图片后共获取到725个有效样本，将每张图片重命名为其所表示的四位字符验证码作为标签。

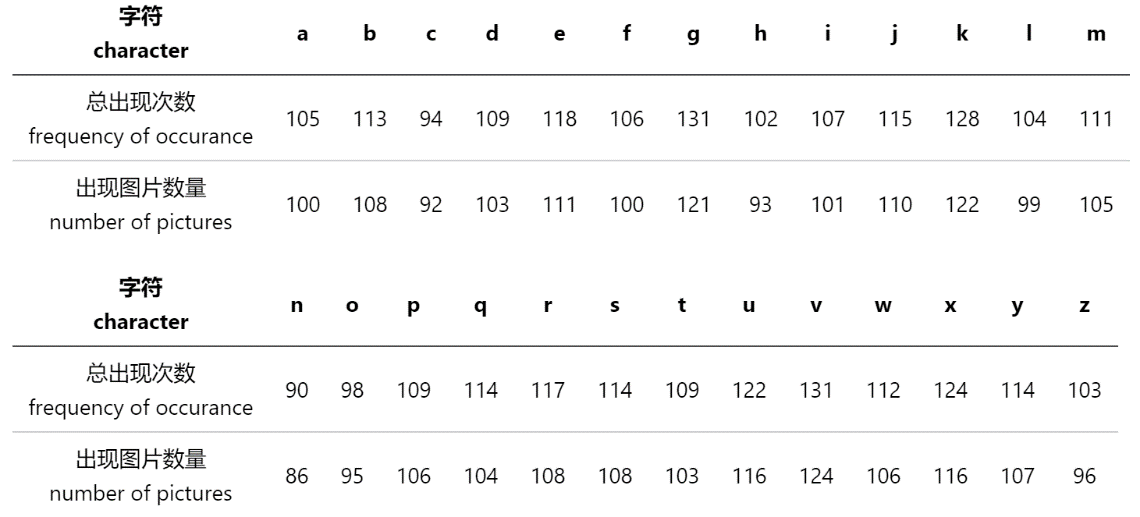
实验表明二值化后的验证码图片识别效果优于原始图片，故将原始样本二值化得到新的数据集。

* 1. 数据特征分析

观察样本，发现样本图片整体较为清晰可辨，目标字符均为英文字母，干扰较少、字符间偶有重叠，具有应用OCR方法的基本条件。由程序统计，725张图片中26个英文字母的分布情况如表1所示，可见各字符分布较为均匀，数据集整体相对均衡，可应用深度学习方法。

表1 数据集字符分布情况

Tab.1 Distribution of characters in dataset



* 1. 数据增强方法

针对图像数量较少难以使卷积神经网络学习到足够多的特征这一问题，使用数据增强来扩充样本。使用augmentor库中的随机扭曲(random\_distortion)方法在原始图像中以0.5的概率在2×2的像素框格中进行扭曲，扭曲效果对比与二值化后效果对比如图2所示。



图2 数据增强效果与二值化效果对比

Fig 2 Comparison of data enhancement effect and binarization effect

augmentor库提供了扭曲、倾斜、旋转和随即干扰等数据增强手段，为保留验证码的原始特征故仅选择小幅度扭曲来增强数据。

由于表示图片真值的标签作为文件名，为避免文件名重复，为文件名添加计数后缀。

* 1. OCR方法

OCR即光学字符识别，指电子设备如扫描仪或数码相机检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。一般包括二值化、噪声去除、倾斜校正、版面分析、字符切割、字符识别和版面恢复等过程。现阶段OCR已较为成熟，许多软件提供效果可观的OCR服务。

为尽快验证可行性，先使用easyocr构建程序原型。

* 1. CNN方法

自20世纪90年代第一个卷积神经网络LeNet-5提出以来[7-9]，CNN在图像分类及识别任务上的性能逐年提升，达到甚至超过了人工水平。本实验所采用的CNN[10]结构如图3所示。1×80×25的图像作为输入，经过四次3×3卷积(Conv)、2×2最大池化(MaxPooling)、批量规范化层(BatchNorm)和ReLU激活函数的结构，数据维度依次变为16×40×12、64×20×6、512×10×3和512×5×1，经过0.5的Dropout进入全连接层(FC)，输出为104×1的向量，每26个为一组表示一个字符的预测值，使用softmax将向量映射为4位字符。

本文实验所用CNN最初版本在FC层前无Dropout层。

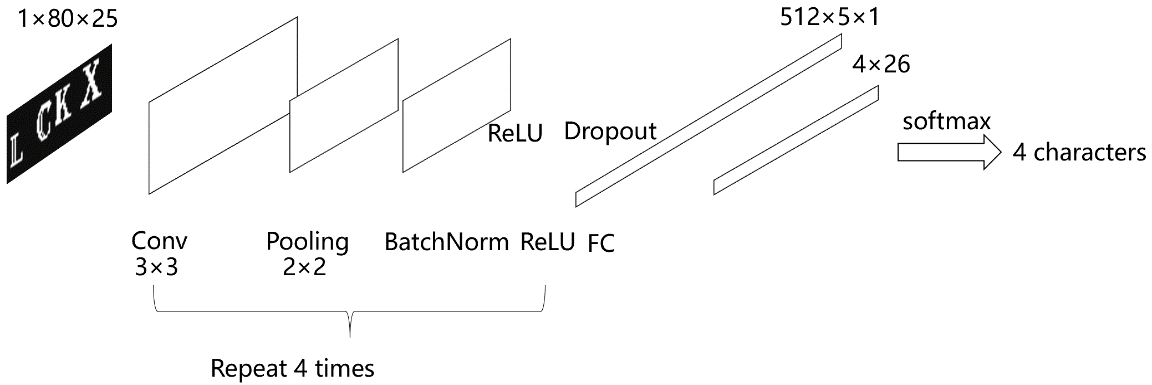


图3 CNN结构

Fig 3 CNN Architecture

许多字符识别任务均先将字符分隔开，再对每个字符进行识别，实际上easyocr就遵循这样的过程，但效果并不理想。因此我们选择深度学习，希望网络可以学习到每个字符的形状特征以及相对次序与输出次序的关系。

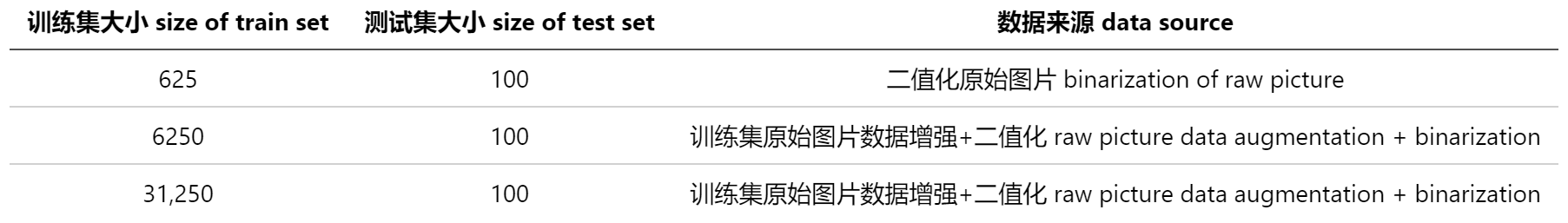
1. 实验结果与分析
   1. 实验设置

本研究的硬件平台CPU为E5-2620，GPU为GTX2080Ti的工作站，操作系统为Ubuntu 20.04.3。Python版本为3.8.12，easyocr版本为1.4.1，Pytorch版本为1.10.0。

实验过程中共使用过以下3种规模的数据集，如表2所示。

表2 实验所用数据集描述

Tab.2 Description of datasets



* 1. 评价指标

本文中验证码识别正确指模型输出的四位字符与图片表示的字符均相同且顺序一致，单字符识别正确指模型输出的字符与图片中对应次序的字符相同。

OCR方法的评价指标为验证码识别的准确率；CNN方法的主要评价指标为验证码识别准确率，以单字符识别准确率作为参考指标。若用accuracy表示准确率，T表示识别正确的样本数量，F为识别错误的样本数量，即有

* 1. OCR方法实验结果及分析

不同策略下easyocr在100张图像上的验证码识别效果如表3所示。仅使用easyocr以英文字符集作为目标字符集对原始验证码图片进行识别，准确率仅为1.98%。实验过程中发现模型对于某些样本可以识别出正确的字母，但在输出的序列中夹杂空格符，显然正确结果一定为4个英文字母表中的字符，因此对easyocr的输出首先进行空格过滤，识别准确率提升至12.87%。此外，为避免原始图片中的色彩及图像噪点干扰，将图像进行二值化处理再作为模型输入，再过滤输出结果中的空格符，准确率提升至28.71%。

分析认为存在以下原因：验证码图片存在字符重叠的特点；字母不严格处于同一行对OCR干扰较大；OCR模型所识别字符集范围过大使得OCR方法准确率较低。

表3 不同策略下easyocr的识别效果

Tab.3 Recognition effect of *easyocr* under different strategy

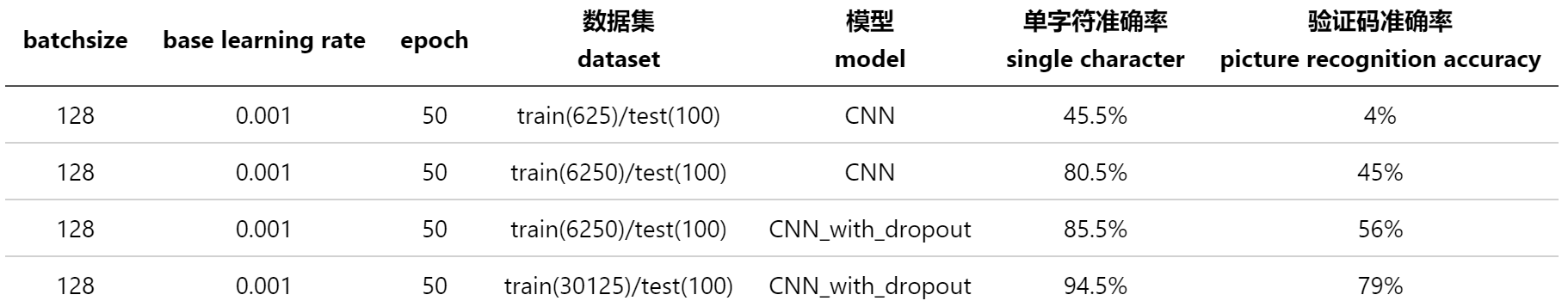


可见OCR的效果最终仍不理想，在总共725张样本的数据集上也仅达到24.69%的准确率，但这并不代表该方法无法工程应用。尤其是当识别登录界面的验证码是一个可重复的过程时，我们可以大量重复这一过程来弥补准确率的不足。单次验证码识别的错误率约为75.31%，则连续对9张随机验证码的识别均失败的可能性约为0.7531^8=0.0779，即9次均失败的概率为7%，换言之，在工程实践中，我们可以认为识别验证码这一过程连续重复9次成功的概率为93%。结合爬虫程序实践表明，OCR模型通常能在10次尝试中输出一个正确结果。

* 1. CNN方法实验结果及分析

表4 CNN方法实验设置及结果

Tab.4 Experimental setup and results of CNN



如表4所示，数据扩充后单字符准确率和验证码识别准确率均显著提高。在CNN全连接层前增加Dropout层后两种准确率也提高了约10%。

此外还有一个有趣的现象，验证码识别准确率约为单字符识别准确率的4次方，即验证码识别任务如同分解为了4次单字符识别任务。若将4个字符作为一个整体看待，这与卷积神经网络的平移不变性并不冲突。平移不变性的位置无关更像是整体位置而非字符间的相对位置。

在与爬虫程序的结合中，由于CNN可观的准确率，验证码识别的尝试次数明显少于OCR，从而使登录所需时间更短。

1. 结论与展望
2. 在可重复尝试的任务中，模型准确率低不代表无法取得良好的工程效果。
3. 对于CET6报名网站的验证码，简单的CNN网络即可达到94.5%的单字符识别准确率和79%的验证码识别准确率。
4. 该网络的泛化能力有待进一步验证。
5. 验证码识别任务是否被等价或近似分解为了4次单字符识别任务有待进一步实验验证。

参 考 文 献

1. 全国大学英语四、六级考试报名网[EB/OL]. https://cet-bm.neea.edu.cn/
2. 北京林业大学关于2021年下半年全国大学英语四、六级考试报名的通知[EB/OL]. http://jwc.bjfu.edu.cn/ksxx/378181.html
3. 西安邮电大学关于2021年12月CET考试报名工作的通知[EB/OL]. http://jyc.xupt.edu.cn/info/1195/2971.htm
4. 汪六三, 黄子良, 王儒敬. 基于近红外光谱和机器学习的大豆种 皮裂纹识别研究[J]. 农业机械学报, 2021, 52(6)：361−368. Wang L S, Huang Z L, Wang R J. Identification of soybean seed coat crack using near infrared spectroscopy and machine learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(6): 361−368.
5. 杨晓花, 高海云. 基于改进贝叶斯的书目自动分类算法[J]. 计算机科学, 2018, 45(8)：203−207. Yang X H, Gao H Y. Improved Bayesian algorithm based automatic classification method for bibliography[J]. Computer Science, 2018, 45(8): 203−207.
6. Majdar R S, Ghassemian H. A probabilistic svm approach for hyperspectral image classification using spectral and texture features[J]. International Journal of Remote Sensing, 2017, 38(15): 4265−4284.
7. Lecun Y . Generalization and Network Design Strategies[C]// Connectionism in Perspective. Elsevier, 1989.
8. Y. LeCun *et al*., Backpropagation Applied to Handwritten Zip Code Recognition Neural Computation[J]. vol. 1, no. 4, pp. 541-551, Dec. 1989.
9. Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
10. 使用Pytorch搭建CNN识别验证码[EB/OL]. https://github.com/ice-tong/pytorch-captcha