基于渐进式学习的神经网络端到端验证码识别

刘达荣,张远平,汤茂斌,李福芳
(广州大学 计算机科学与教育软件学院,广东 广州 510006)

摘 要:针对验证码经过弯曲变形,无法采用传统的字符分割方法进行检测的问题,在模仿人类的渐进式学习过程的基础上,提出利用卷积神经网络,优化手写识别 MNIST 的三层网络结构,无需预先对验证码进行分割,直接对验证码进行端到端的识别。利用不确定度回收训练图片,减少训练集数据量,并利用可视化工具提高其网络识别性能。经过 55 万次训练后,生成了检测模型并对测试集验证码进行了检测,验证码识别速度达到 0.073 秒/张,准确率达到 86%。通过对比同一测试环境下的测试集,发现利用渐进式学习方法具有更高的建模效率和更好的识别准确率,并对识别错误的验证码进行分析。

关键词:渐进式学习;卷积神经网络;验证码;无分割;端到端
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2018)09-0016-04
doi:10.3969/j.issn.1673-629X.2018.09.004

End-to-end Verification Code Identification of Neural Network Based on Progressive Learning

LIU Da-rong, ZHANG Yuan-ping, TANG Mao-bin, LI Fu-fang
(School of Computer and Education Software, Guangzhou University, Guangzhou 510006, China)

Abstract: For authentication code can't be measured by traditional character segmentation method after bending deformation, on the basis of the progressive learning of imitating human, we put forward convolution neural network to optimize three layers network structure of handwritten recognition MNIST, without prior segmentation for authentication code, directly identifying in the end-to-end to authentication code. The training images are recovered according to uncertainty to reduce the amount of training data, and the network recognition performance is improved with visual tools. After training by 550 000 times, the detection model is generated and the verification code from the test set is detected. The recognition speed of the verification code reaches 0.073 seconds per piece, and the accuracy is 86%. By contrast with the test set in the same test environment, it is found that the progressive learning method has higher modeling efficiency and better recognition accuracy. The wrong verification code is analyzed.

Key words: progressive learning; convolution neural network; verification code; no segmentation; end to end

1 概 述

验证码[1]作为区分人类和机器的测试工具,在图灵测试上占有一席之地。如铁道部在高铁票购买过程中,进行验证码识别以防止黄牛党抢票;论坛注册时,提供的验证码测试防止脚本自动注册。目前,比较常用的验证码都会包含一些扭曲变形的字符进行测试,测试者必须能看懂并输入正确的字符串才能通过。对于人类来说,扭曲变形的字符并不难识别,但对于机器就相当困难,一个好的验证码可以有 80% 的概率被人类识别[2],但对于机器只有 0.01% 的概率被识别。一个典型的验证码样例[3]如图 1 所示。

图 1 谷歌提供的验证码样例
(包含一串扭曲变形的字符)

传统的光学字符识别[4]的方法是:首先定位字符

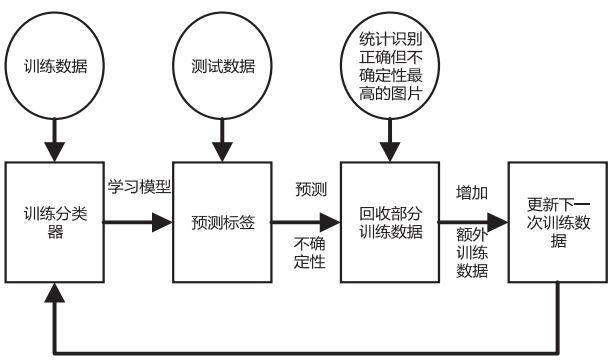
收稿日期:2017-08-24 **修回日期**:2017-12-27 **网络出版时间**:2018-05-16
基金项目:国家自然科学基金(61472092)
作者简介:刘达荣(1988-),男,硕士研究生,研究方向为图像处理、深度学习;张远平,教授,研究方向为计算机算法设计与分析;汤茂斌,副教授,研究方向为人工智能、数据挖掘、系统分析与设计;李福芳,副教授,研究方向为计算机网络、网络(分布式)计算。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180515.1645.006.html

的位置,然后进行分割[5]和识别。Mori 等[6-7]提出先分割,后使用字典来解决验证码识别问题。然而,如图 1 所示,因为这种验证码的字符发生重叠或者变形,已经不能用垂直线分割。目前,相关工作者已经开始利用人工智能[8]研究自动化方法去解决验证码识别问题,如 Goodfellow 等[9]使用深度卷积神经网络(CNN)成功解决了该问题。但是训练一个深度卷积神经网络需要非常大的数据集[10],如 ImageNet[11]数据集需要大量的图片数据:训练数据集 126 万张图像,验证集 5 万张,测试集 10 万张。对于验证码识别,具有人工标签的训练数据并不多,而且与自然图片中的对象不同,验证码的变化空间非常大。

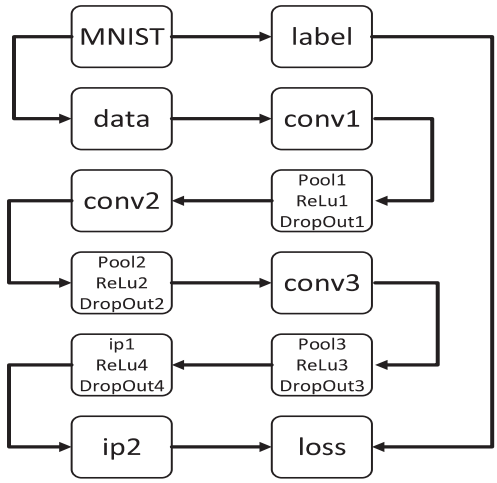
由于变形验证码的特征与手写字符类似,因此文中借鉴 LeCun 等[12]提出的卷积神经网络架构进行识别。CNN 被设计成层级结构,常应用于目标检测和分类,由于带有卷积层,对图像处理具有特别优势。利用卷积对图像特有的局部感受野进行特征提取,并进行权值共享,从而大幅减少参数;利用多卷积核学习多种特征,在通过卷积获得了特征之后,利用池化对不同位置的特征进行聚合统计[13]。目前,基于 CNN 产生的模型基本能解决大部分视觉问题,可见其强大功能。

训练一个深度神经网络需要较大的数据集和耗费较长的训练时间,对此,文中提出两种新方法:使用“best-vs-second-best”统计方法对验证码进行回收,以减少训练数据量;模仿人类的学习模式—渐进式学习,由易到难进行学习来减少训练时间,提出基于渐进式学习的卷积神经网络,对验证码进行端到端的识别。

实验流程如图 2(a)所示。首先,利用一万张没有扭曲变形的验证码图片作为训练数据,基于手写识别 MNIST 优化后的网络进行预训练,训练完毕后再对十万张验证码图片进行预测,在输出结果时利用“best-vs-second-best”统计方法,把不确定度较大的图片放回下一次训练中;接着,当正确率上升到比较高的时候(90% 以上),把原来的验证码图片替换成扭曲变形较大的图片,重复以上过程,从而缩短学习时间并获取较高的准确率。



(a)训练流程和数据流向



(b)训练分类器的网络结构

图 2　实验流程以及训练分类器的网络结构

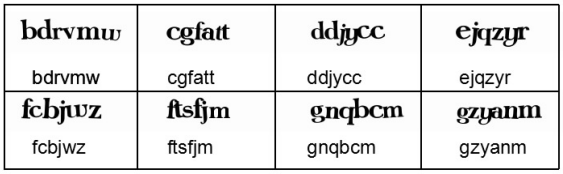
2　学习模型

传统的神经网络激活函数以 Sigmoid 函数和 Tanh 函数为主[14]。从数学上来看,非线性的 Sigmoid 函数对中央区的信号增益较大,对两侧区的信号增益小,在信号的特征空间映射上有很好的效果。从神经科学上来看,中央区酷似神经元的兴奋态,两侧区酷似神经元的抑制态,因而在神经网络学习方面,可以将重点特征推向中央区,将非重点特征推向两侧区。而在深层网络中,为了加快训练网络和减少梯度,利用了 ReLu(rectified linear units)激活函数[15]。

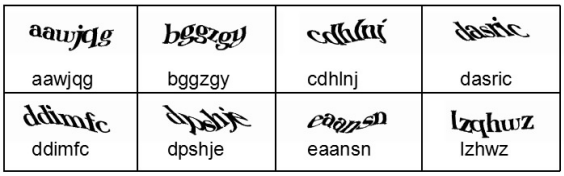
基于手写识别 MNIST 的三层网络结构,文中对其进行改造,在池化层加入 ReLu 激活函数来提高收敛速度,以及加入 DropOut 技术[16-17]来防止网络过拟合,如图 2(b)。DropOut 技术是在训练网络时用的一种技巧,在训练开始时,随机选择一半隐层单元,保持输入和输出层不变进行权值更新,在第二次迭代中,再次随机选择一半隐层单元,以此类推直到训练结束。运用 DropOut 的训练过程,相当于训练了很多个只有半数隐层单元的神经网络,每一个这样的半数网络,都可以给出一个分类结果,这些结果有的是正确的,有的是错误的。随着训练的进行,大部分半数网络都可以给出正确的分类结果,那么,少数的错误分类结果就不会对最终结果造成较大影响。

2.1　训练数据的生成

由于没有手动打标签的验证码数据集,文中采用自己编写的 PHP 脚本自动生成验证码,并保证没有重复的验证码图片。使用 Cool PHP CAPTHCAH 框架生成验证码,其中包含了一串固定长度(6 位)的扭曲字符,类似于谷歌的 reCAPTCHA[4],生成的验证码图片大小为 50×180。调整该框架生成灰度图,没有加入阴影和横线作为背景,部分验证码数据集见图 3。



(a)没有扭曲变形的简单验证码



(b)经过扭曲变形的困难验证码

图 3 验证码

2.2 卷积层网络结构参数和特征可视化

如图 2(b)所示,图片以像素形式转换成矩阵输入网络。在卷积层 1 中,对数据层以窗口大小 5×5、步长为 2 进行特征提取,把原图片 50×180 压缩编码以 25×90 的矩阵向量输出到下一层,直到最后的分类输出层。在卷积层 1、2、3 的输出中,第一个输出图学习到了验证码的边缘,第二个输出图学习到了验证码的局部特征,第三个输出图学习到了验证码的全局特征。

基于 Caffe 框架[18],硬件采用 GTX960 GPU 进行训练,基础学习率(base_lr)初始值设置为 0.01,动量为 0.9,权值衰减为 0.000 5,学习率策略采用多项式,权重(power)为 0.75,最大迭代次数(max_iter)为 20 000,并使用 GPU 进行加速训练,迭代次数(iter)为两千次。由于选取学习率策略为多项式模式(Poly),故在训练过程中的有效学习率为 base_lr * (1-iter/max_iter)^(power),训练两万张图片。

对超参数进行设置和建模,以及对每一层的输入与输出进行特征可视化,如图 4 所示。

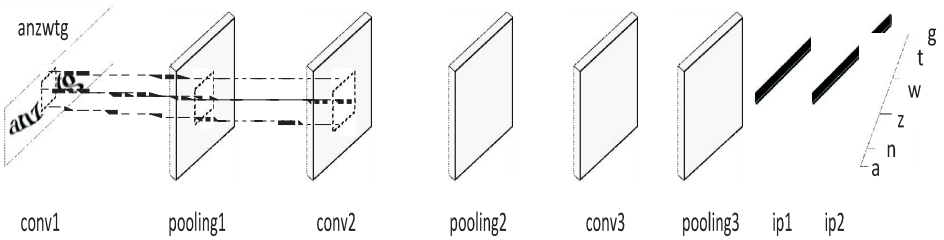


图 4 网络结构

2.3 输出层

对于输出层,为预测每个验证码图片的字符概率,采用 SoftmaxWithLoss 多分类函数:

$$P(y_i \mid v) = \frac{e^{v^T\omega_i}}{\sum_{k=1}^{K} e^{v^T\omega_k}} \tag{1}$$

即对于隐层的输出 v,第 i 个输出单元 y_i 的概率为 $P(y_i \mid v)$,ω 为隐层与输出层的权重值。在输出层,由于不能直接输出字符,采用如下映射函数:

$$\Theta(y_i) = \begin{cases} '0'\cdots'9', y_i = 1 + 62i, 2 + 62i, \cdots, 10 + 62i \\ 'A'\cdots'Z', y_i = 11 + 62i, \cdots, 36 + 62i \\ 'a'\cdots'z', y_i = 37 + 62i, \cdots, 62 + 62i \end{cases} \tag{2}$$

其中,y_i 为输出层第 i(i =0,1,…,5)个单元,对应从数字 0 到字母 z 的其中一个,其余如此类推。如第一个输出单元 y_0,输出最大概率对应的横坐标为 37,代表输出字符 a;第二个输出单元 y_1,输出最大概率对应的横坐标为 112,代表输出字符 n,如此类推,如图 5 所示。

以上文的验证码图片“anzwtg”为例,图中为六位字符经过识别后每一位对应的最大概率,即 x 轴的第 0 号到 61 号为第一个字符的分布区间,第 62 到 123 为第二个字符的分布区间,依此类推。

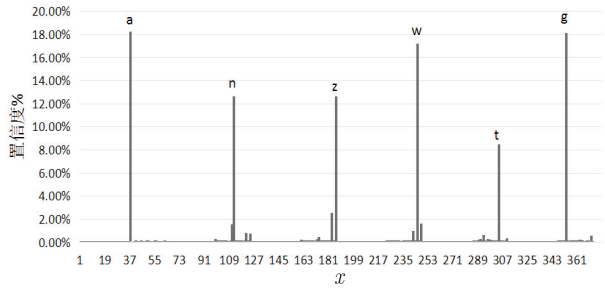


图 5 提取输出层的概率分布图

2.4 利用不确定度回收训练图片

文中采用“best-vs-second-best”来统计已经分类正确但具有较大不确定性的图片,重新放回下一次迭代训练,以此减少训练样本大小,公式如下:

$$\eta = \frac{1}{d} \cdot \sum_{i=1}^{d} \frac{\operatorname{argmax}\{P(y_i) \setminus \operatorname{argmax} P(y_i)\}}{\operatorname{argmax} P(y_i)} \tag{3}$$

其中,d 为迭代次数;$P(y_i)$ 为每个字符的概率大小。分母是每次迭代的字符最大概率,分子是每次迭代的字符次大概率,即选取总不确定度 η 值最大的图片,重新放入下一次迭代训练,以此来减少训练图片规模。

3 实验结果与分析

文中提出渐进式学习,其准确率和学习过程如图 6 所示。其中,实线为直接采用困难验证码图片训练

的效果,虚线为先进行简单图片预训练,当准确率达到 98% 时,替换成困难验证码图片进行训练的效果。

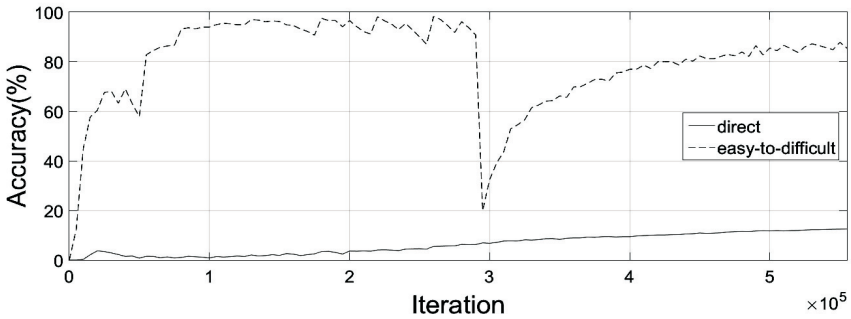


图 6　迭代次数与准确率

从图 6 可以明显看出,采用渐进式学习在同一训练时间内准确率更高,达到 86%。本次实验中,训练的总时间约为 48 h,若继续增加训练时间,准确率还可以更高。与传统方法相比较,在单张验证码识别时间上,基于 SVM 识别需要 0.76 s[19],在本测试中需要 0.073 s,识别效率约为传统方法的 10 倍。

4　结束语

结合传统神经网络的结构进行优化,采用模仿人类学习过程的渐进式学习模式,提出一种无分割、端到端验证码识别算法。该算法充分利用卷积神经网络的自学习特性,简化传统验证码识别中字符分割等人工干预手段,使得网络具有抗旋转的优良特性,并利用概率统计方法减少训练样本,最终使得神经网络具有收敛速度快、检测效果好的特性。在后续工作中,将在背景加入更多干扰来进行测试,以期提高网络的鲁棒性。

参考文献:

[1] VON AHN L, BLUM M, HOPPER N J, et al. CAPTCHA: using hard AI problems for security[M]//Advances in cryptology. Berlin: Springer, 2003: 294-311.

[2] CHELLAPILLA K, SIMARD P Y. Using machine learning to break visual human interaction proofs (HIPs)[C]//Advances in neural information processing systems. [s. l.]: [s. n.], 2004: 265-272.

[3] AHN L V, MAURER B, MCMILLEN C, et al. reCAPTCHA: human-based character recognition via web security measures[J]. Science, 2008, 321(5895): 1465-1468.

[4] JADERBERG M, VEDALDI A, ZISSERMAN A. Deep features for text spotting[C]//European conference on computer vision. Berlin: Springer, 2014: 512-528.

[5] LU Yi. Machine printed character segmentation: an overview[J]. Pattern Recognition, 1995, 28(1): 67-80.

[6] 吕　刚, 郝　平. 基于神经网络的数字验证码识别研究[J]. 浙江工业大学学报, 2010, 38(4): 433-436.

[7] 左保河, 石晓爱, 谢芳勇, 等. 基于神经网络的网络验证码识别研究[J]. 计算机工程与科学, 2009, 31(12): 20-22.

[8] 邓介一, 刘黎志, 谭培祥. 基于神经网络的数字字符识别系统设计与实现[J]. 软件导刊, 2017, 16(5): 47-50.

[9] GOODFELLOW I J, BULATOV Y, IBARZ J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks[C]//Computer vision and pattern recognition. [s. l.]: [s. n.], 2013.

[10] 付先珺. 基于 RPROP 人工神经网络对验证码识别的研究与实现[D]. 重庆: 重庆大学, 2011.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th international conference on neural information processing systems. Lake Tahoe, Nevada Curran: Associates Inc., 2012: 1097-1105.

[12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[13] 高　原. 基于 BP 神经网络的文本验证码破解[J]. 电子科技, 2012, 25(7): 37-42.

[14] 吕　霁. 基于神经网络的验证码识别技术研究[D]. 泉州: 华侨大学, 2015.

[15] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th international conference on machine learning. Haifa, Israel: Omnipress, 2010: 807-814.

[16] 陈　超, 毛坚桓, 刘　寅. 基于卷积神经网络的铁路货运网站验证码识别[J]. 指挥信息系统与技术, 2016, 7(4): 91-96.

[17] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[18] JIA Yangqing, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]//International conference on multimedia. [s. l.]: ACM, 2014: 675-678.

[19] 唐海涛. 自组织增量神经网络的验证码识别模型与算法[D]. 广州: 广东工业大学, 2016.