引用格式: 沈强,李辉,张燕. 基于小规模手写体汉字数据集的数据增强方法[J]. 北京化工大学学报(自然科学版) ,2021,48(1) :58 -65.
SHEN Qiang, LI Hui, ZHANG Yan. A data augmentation method based on a small-scale handwritten Chinese character dataset[J]. Journal of Beijing University of Chemical Technology (Natural Science) , 2021,48(1) :58 -65.

基于小规模手写体汉字数据集的数据增强方法

沈　强[1]　李　辉[1*]　张　燕[2]
(1. 北京化工大学 信息科学与技术学院, 北京　100029; 2. 北京市电气工程学校, 北京　100123)

摘　要: 针对深度卷积生成对抗网络(DCGAN) 在小规模手写体汉字数据集下生成数据重复多样、分类效果较差的问题,提出结合传统数据增强方法的结合式生成方法 X-DCGAN。该方法通过预增强模块给予神经网络部分更充足多样的训练数据,减少因网络过拟合与训练不充分而出现的样本重复率高、学习效果较差的状况。实验结果表明,本文方法生成的样本数据较单一方法在样本多样性方面显著提高,生成数据进行分类测试时获得的平均识别率较 DCGAN 方法提升了 9.67%。X-DCGAN 充分发挥了传统数据增强方法和生成式方法各自的优势,能够更加有效地解决小规模数据集的扩展与增强问题。

关键词: 数据增强; 深度卷积生成对抗网络; 手写体汉字识别; 图像处理
中图分类号: TP391.4　　**DOI**: 10.13543/j.bhxbzr.2021.01.008

引　言

手写体汉字识别长期以来都是模式识别的一个重要研究领域,由于汉字字符种类庞大、书写风格迥异、存在很多形似字符以及手写体汉字数据集不足等原因,给识别带来极大挑战。近年来,基于卷积神经网络(CNN) 的图像识别方法取得了大量成果。Krizhevsky 等[1] 在 2012 年 ImageNet 竞赛上提出的 AlexNet 网络模型以超过第二名 10.9% 的绝对优势一举夺冠,并由此在该领域内引发 CNN 研究热潮; 2015 年,GoogleNet 模型在比赛中脱颖而出[2],在加强网络深度的同时提出了创新结构; 2016 年,He 等[3] 提出 ResNet 网络结构,该结构拥有超过 100 层的网络,通过引入残差单元来解决退化问题,成为里程碑事件。在手写体汉字识别方面也有诸多学者提出不同特点的训练网络,均获得了不错的分类效果[4-6]。在以上文献中,除改进模型结构外,良好性能的取得也离不开大规模数据集的支撑。原因在于卷积神经网络需要训练大量带有标签的数据样本才能获得更优的分类效果,小规模数据集训练容易出现过拟合现象,即神经网络模型过于复杂(例如 AlexNet 具有 6 千万级的参数规模) 虽然可以很好地适应每一个训练数据的分布,但会忽略其中的通用趋势,使之对未知数据无法作出可靠的判断。

目前研究领域内常见的手写体汉字数据集存在字符类别少、书写者数量少(字符多样性差) 、样本不均衡和总体样本数少等不足。例如北京邮电大学发布的 HCL2000[7] 与哈尔滨工业大学发布的 HIT-MW 手写体数据集[8],两者虽采样于真实书写环境,具有丰富多样的书写风格,但字符种类和样本规模较小; SCUT-COUCH2009 是由华南理工大学发布的具有最多字符种类的手写体字符集[9],共计 13 548 类汉字,数据规模达到 360 万; 相较而言,由中国科学院发布的 CASIA-OLHWDB/HWDB 数据集拥有最大的数据规模—520 万脱机手写体字符样本[10]; 目前拥有最多书写者的手写体汉字数据集是包含 2 988 名作者样本的 SCUT-EPT[11],样本共涵盖 4 250 类汉字,规模为 120 万。除数据规模外,在数据集建立过程中还需要设计合理的条件限制以及投入大量的时间成本和人工成本,这些也是阻碍大规模数据集建立的重要因素。基于现实条件的限制,很多时候无法获取到足够的训练数据来解决特殊环境下的字符识别问题,例如医疗文本、刑侦文本、少

收稿日期: 2020-03-31
第一作者: 男,1994 年生,硕士生
* 通信联系人
E-mail: ray@mail.buct.edu.cn

数民族文字等等,这些数据集规模远小于汉字数据集,并且建立难度更高。

为了解决样本数据不足的问题,人们最先提出了基于图像处理的数据增强技术方法。例如,有学者提出利用平移、尺度缩放、变形拉伸、旋转等方法对汉字图像数据进行扩充并验证了方法的有效性[12]。由 Simard 等[13]提出的弹性形变(elastic distortion) 方法在 Mixed National Institute of Standards and Technology Database(MNIST) 上大获成功。随后数据增强技术在字符识别领域得到广泛应用[14-16]。生成对抗网络(generative adversarial networks,GAN) 则是 Goodfellow 等[17]在 2014 年提出的一种基于博弈论思想的生成式网络模型,作者从理论上证明了 GAN 模型可以生成与真实数据相同分布的数据。之后,以 GAN 为核心的数据增强方法开始被不断提出[18-20]。至此,数据增强技术大致可分为两类,即传统方法(包括形变类、噪声类) 和生成式方法。形变方法使得字符形状发生变化,通过造成字符的结构信息改变来模拟不同人群的书写习惯和书写风格。噪声方法增加或改变图片背景信息,对字符本身结构不产生直接影响,模拟的是书写环境造成的冗余信息。生成式方法更倾向于模仿,神经网络通过学习训练生成能够代替真实样本的伪数据。

目前生成式方法在连续性图形类数据中应用较为广泛,由于离散型文本数据较难学习,再加上网络本身需要一定数量的训练数据,使得单一使用生成网络方法对小规模数据的字符数据增强效果并不显著。基于小规模手写体汉字数据集,本文提出一种结合传统数据增强技术的深度卷积对抗生成网络方法(X-DCGAN) 进行数据集扩充和识别,该方法能够结合两者优势,给予生成网络更充足的训练数据,从而显著提高生成式数据增强方法在小规模手写体汉字数据集下的生成效果。

1　传统数据增强方法

常见的传统数据增强方法为形变类和噪声类。形变类改变字符形态,如笔画粗细、字符位置、笔画扭曲等,使得字符在原有形态上呈现出多种书写变化,如图 1 所示。噪声类对图片整体进行处理,增加或改变像素信息,达到模糊、杂乱及离散化效果,如图 2 所示。

1.1　形变类方法

形态学操作(Thickness)　该方法包括几类基本的图像运算,如膨胀(Dilate) 、腐蚀(Erode) 、开运算、闭运算和形态学梯度等。这些操作能够改变字符笔画的宽度,可用于图像的边界提取、区域填充、细化、粗化等。

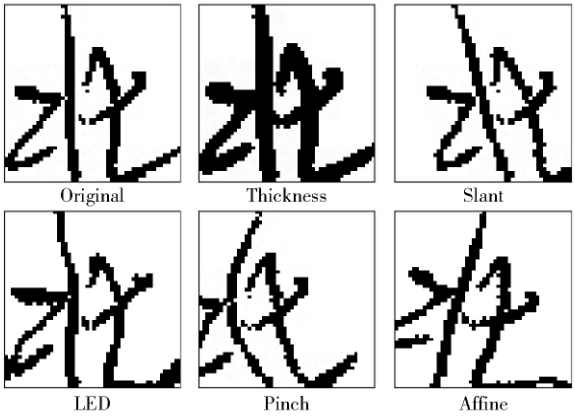

图 1　形变类数据增强方法效果

Fig. 1　Effect of deformation method

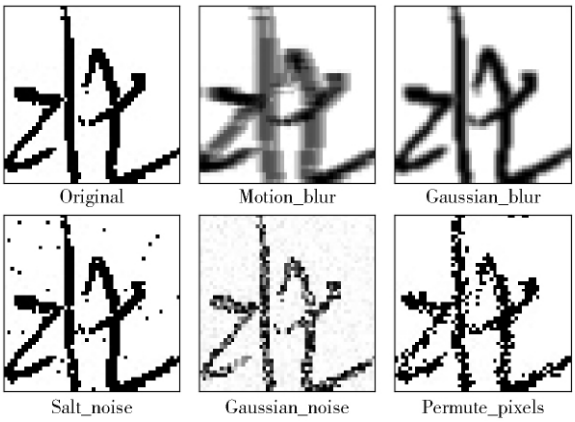

图 2　噪声类数据增强方法效果

Fig. 2　Effect of noise method

图像倾斜(Slant) 和仿射变换(Affine)[21]　这两种方法能够使得图像产生角度变化,达到放大、缩小、平移和旋转的效果,并且允许图像在两个方向上实现任意的伸缩与变换。

凹凸变换(Pinch) 和局部弹性变换(local elastic deformations, LED)[22]　它们通过对图像进行卷积计算,能够造成不规则的变化,从而模拟真实书写过程中笔画扭曲的效果。

1.2　噪声类方法

运动模糊(Motion_blur)　该方法通过一维卷积操作使得图像呈现出物体运动的效果。

高斯模糊(Gaussian_blur) 和高斯噪声(Gaussian_noise)　前者是一种低通滤波器,能够用于消除图

像的噪声,尤其对高斯噪声有明显作用,但两种方法均能在图像中形成噪声。

椒盐噪声(Salt_noise) 和像素交换(Permute_pixels) 这两种方法能够模拟图像信息传输过程中的信号干扰效果,在图片背景上随机生成白色(或黑色) 噪点,从而形成不同的噪声效果。

2 生成式数据增强方法

2.1 生成对抗网络

深度学习的发展是为了发现丰富的层次模型,即人工智能应用中数据的概率分布如自然图像、语音、语言符号等。由于使用最大似然估计和相关策略难以描述这样的概率计算,导致已有的深度生成模型学习效果一般。为了更好地解决这类问题,Goodfellow 等[17] 提出了新的生成模型估计方法,即生成对抗网络。

GAN 框架中会同时训练一个生成模型 G 用以捕获样本数据分布并模仿生成伪数据分布,和一个判别模型 D 用以判断样本是来源于训练数据或者 G。理想情况下,在任意函数空间 G 和 D 中存在唯一解,使得 G 生成的数据分布由 D 判断正确的概率为 1/2。定义真实训练数据为 x 及其分布为 P_{data},将随机噪声 z(其分布表示为 $P_z(z)$) 输入 G 进行学习训练,生成的数据分布为 P_g,则噪声变量在数据空间的映射表示为 $G(z;\theta_g)$,其中 G 是具有 θ_g 参数的多层感知器表示的可微函数。同理,定义判别器模型 $D(x;\theta_d)$ 表示输入数据的来源,其输出的是判别结果为真实训练数据的概率值。G 的目标是尽可能生成符合真实训练数据的分布,D 的目标是尽可能准确地判断数据来源。使用值函数 $V(G,D)$ 表示两者之间极大极小的博弈局面,如式(1) 所示。

$$\min_G \max_D V(G,D) = E_{x\sim P_{\text{data}}(x)}[\lg D(x)] + E_{z\sim P_z(z)}[\lg(1-D(G(z)))] \quad (1)$$

根据式(1) 的描述,在不增加优化条件的情况下,实践中会存在两个问题: 一是在学习早期 D 可以高置信度地辨别样本,因此 G 获得的梯度效果十分有限; 二是在有限数据训练的情况下,D 优化结束后将会导致过拟合,使得模型无法收敛。以上两个问题均会导致整体模型不稳定,出现博弈结果单一化的情况。故而我们提出随机梯度下降算法,其中要求 G 和 D 的优化交替进行,并且首先通过提升随机梯度优化 D,这个过程需要运行 k 次(k 是一个超参数) ,然后再通过降低随机梯度更新 G。

原始 GAN 方法的主要缺陷显而易见,即模型一次只能学习一类数据从而导致生成效率低下,但通过反向传播方式获得梯度、在学习过程中可以将多种函数合并到模型等优势,形成更强的图像生成能力,并且促进了基于 GAN 的新型生成式网络的发展。

2.2 深度卷积生成对抗网络

使用 GAN 进行数据增强的目的在于扩充训练数据的规模和样本多样性,但不稳定的网络结构导致生成器经常输出无意义图。Horsley 等[23] 描述了一种具有一定结构约束的深度卷积生成对抗网络(DCGAN) ,DCGAN 作为一种有力的无监督学习方法可以从大量未标记数据集中学习到可重用特征表示,相较于传统 GAN 具有更强大的生成能力,有效地解决了网络的不稳定问题。文献[23]的研究结果表明 DCGAN 在大多数情况下均能稳定地训练,完成训练的判别器 D 还可以用于分类,并且效果优于其他无监督学习,另外 G 具有的向量算术特性使得生成的样本丰富多样。

深度卷积生成对抗网络的判别器和生成器均是卷积神经网络结构(表 1) 。在判别器 D 中,输入图像经过若干层卷积后获得卷积特征,通过 *Logistic* 函数输出判别结果概率。以生成器的网络结构为例,从输入到输出其实是数据张量变化的过程。如果输入 100 维向量 z,首先会经过一个全连接层将其转换为 4 × 4 × 1 024 维向量,之后使用转置卷积作上采样操作,最后得到 64 × 64 × 3 的图像。为了稳定 GAN,在 DCGAN 中进行的主要变化包括: 1) 使用带步长的卷积替换原始卷积神经网络中的全部池化层; 2) 在生成器和判别器中均使用 *Batch Normalization* 帮助模型收敛; 3) 移除全连接的隐藏层以获得更深的架构; 4) 在 G 中,除了最后一层因需要输出

表 1 DCGAN 各层节点及激活函数

Table 1 Nodes and activation functions of DCGAN

层数	生成器结构		判别器结构	
	节点	激活函数	节点	激活函数
输入	100 × 1	*ReLU*	64 × 64 × 3	*Leaky ReLU*
1	4 × 4 × 1 024	*ReLU*	32 × 32 × 64	*Leaky ReLU*
2	8 × 8 × 512	*ReLU*	16 × 16 × 128	*Leaky ReLU*
3	16 × 16 × 256	*ReLU*	8 × 8 × 256	*Leaky ReLU*
4	32 × 32 × 128	*Tanh*	4 × 4 × 512	*Sigmoid*
输出	64 × 64 × 3		True/False	

图像采用 *Tanh* 函数外其余激活函数均采用 *ReLU* 函数; 5) 在 *D* 中,激活函数主要采用 *Leaky ReLU* 函数。理论上使用 DCGAN 可以获得更加稳定的生成数据。

2.3　新的结合式生成方法

2.3.1　X-DCGAN

本文提出一种结合传统数据增强技术的深度卷积生成对抗网络方法 X-DCGAN,其中“X”表示多种传统数据增强算法的随机选择过程,可以通过数据预增强手段来提升后一阶段的生成效果。在小规模数据集环境下,单一使用 DCGAN 进行数据增强存在诸多问题: 一是 DCGAN 中生成器和判别器的网络结构均采用卷积神经网络,在训练数据不足时易出现过拟合情况,导致生成样本与训练样本关联性强而缺乏变化,无法实现数据多样性扩充的目的; 二是网络对于文本类离散型数据很难进行学习,也使得生成效果欠佳; 三是生成网络计算复杂,作为数据增强器而言略显笨重,并且训练过程时间成本较高。如果通过传统数据增强方式对数据集先进行预扩充操作,再将其输入深度卷积生成对抗网络训练,可以在一定程度上保障网络训练所需的数据规模,充分发挥其生成特性。

传统数据增强方法主要通过图像处理过程改变原始图像的像素分布,以达到变形或变换的目的,虽然已经被多次证明能够有效进行数据增强,但也存在不可避免的缺陷: 一是与原始样本关联性强、变化风格有限; 二是在实践中需要人工设置参数范围,无法完成自动化适应和调整; 三是每次过程都需要进行重复计算,占用大量存储空间,生成效率低下,对大规模数据和图形类数据并不十分适用。传统方法与生成式方法的简单对比见表 2。

表 2　传统方法与生成式方法的优劣比较

Table 2　Comparison of advantages and disadvantages of traditional methods and generative methods

方法	优势	劣势
传统方法	适用小规模数据、对文本数据效果较好	依赖人工操作、模型不可持续化、多样性一般
生成式方法	生成效率高、模型可持续化	需要足够的数据支撑、训练耗费时间成本大

在多类别字符研究缺乏数据支撑时,需要一种能够通过小规模数据即可获得足够数据规模的生成方法,要求生成数据多样并且模型可持续化,同时支持海量数据扩充。综上所述,本文将传统方法和生成式方法的优势进行互补,提出兼具强大生成能力与有效数据增强能力的结合式方法 X-DCGAN。

根据以上思想,X-DCGAN 方法流程设计如图 3 所示,实现步骤如下: 1) 准备小规模数据集,划分为训练集和测试集两个部分; 2) 将训练集数据分组输入预扩充模块,此模块中包含 5 种形变类和 5 种噪声类方法,每组数据进行 k 次随机选择运行数据增强计算,每次计算结果作为一次扩充; 3) 将步骤 2) 中获得的预扩充集作为输入,利用 DCGAN 进行生成过程,将输出结果作为扩充集; 4) 使用预扩充集作为训练数据输入 CNN 分类器,将训练好的模型对预扩充集进行粗分类,用以对 DCGAN 生成的数据标记样本标签; 5) 将扩充集与原始集合并形成最终的增强训练集。

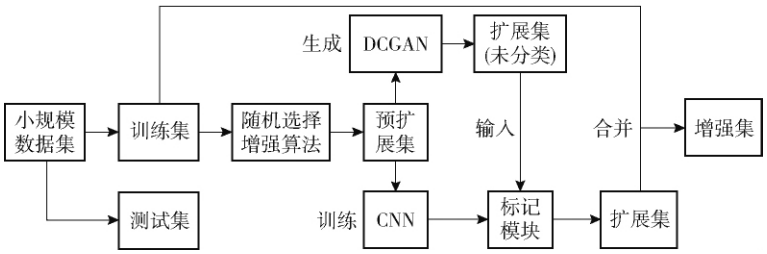

图 3　X-DCGAN 方法流程图

Fig. 3　Flow diagram of X-DCGAN

2.3.2　模块设计

1) 小规模数据集模块设计。本文通过随机问卷收集建立了一个小规模手写体汉字数据集,包含 6 类汉字,564 个真实书写环境下的样本图片(尺寸为 50×50) ,137 位书写者参与。将其作为实验原始样本数据,按照约 30% 的比例进行数据划分,训练集样本个数为 444,测试集样本数为 120。为了比对真实分类效果以及探究在不同环境下采集样本的测试情况,增加 CASIA-HWDB1. 1 数据集中测试集部分对应的 6 类汉字作为补充测试集,样本数为 359。

2) 预扩展模块设计。将小规模数据集的训练集部分作为输入,模块将对每一个样本遍历并进行扩充处理。在 10 类传统数据增强算法中随机选择 k 次(k 表示以单个样本为基础进行图形处理生成的图片个数) 计算,处理后的样本作为预扩展集数据。算法中包括 5 类形变效果算法(膨胀、仿射变换、字符倾斜、凹凸变换和局部弹性变换) 和 5 类添加噪声算法(运动模糊、高斯模糊、椒盐噪声、高斯噪声和像素交换) ,具体模块流程参考图 4。

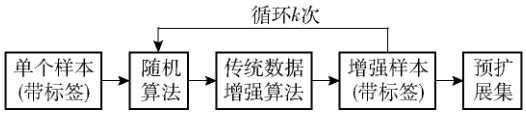

图 4 预处理模块流程图

Fig. 4 Flow diagram of the preprocessing module

在实验过程中,将 444 个小规模样本训练数据依次输入,设置 k 为 9,即对每张样本作用 9 次,使得训练数据规模扩大为 3 996,与原训练集合并后包含 4 440 张样本,图片尺寸不变。输入标记模块和 DCGAN 模块分别训练,称为预扩展集。

3) DCGAN 模块设计。实验中使用的 DCGAN 训练过程主要参数设置如下: 输入图像 50 × 50,进行简单图像处理以适应网络对图像的要求; 输出边长为 *batch_size* × 50 大小的正方形图片,由 *batch_size* 个生成图片组成(生成效率提升的关键因素) ,这里 *batch_size* 设置为 64; 学习率为 0. 000 2,*Adam* 优化器动量为 0. 5; 训练次数设置为 2 000,最终生成样本数量达到 44 400,数据扩充了 100 倍。

4) CNN 分类器模块设计。从第一个成功应用于字符识别问题的卷积神经网络 LeNet 开始,经过不断改进和演化产生了许多优秀的识别模型,如 AlexNet、VGG19、GoogleNet、ResNet 等。随着模型网络层次的逐渐加深和结构的逐渐复杂,模型在分类任务上的表现也在不断提升。在小规模数据集识别实验中,为了排除识别器性能的影响仅考虑数据增强方法效果,设计一个简单 CNN 网络结构,如表 3 所示,所有池化层使用全 0 填充,卷积层不使用。

5) 标记模块设计。基本结构同分类器模块,将训练好的模型载入 CNN 网络中,不进行任何训练步骤,直接执行分类运算的操作。输出分类结果矩阵,同标签分类矩阵进行匹配即可获得输入图像数据的标签序列。

表 3 简单 CNN 结构及各层节点情况

Table 3 Structure of simple CNN and nodes of each layer

层数	详细结构	输入/输出节点数
1	卷积层: 3 × 3 卷积核,6 通道	50 × 50 × 1 /48 × 48 × 6
2	非线性激活函数 *ReLU*	
3	池化层: 步长为 2 × 2	48 × 48 × 6 /24 × 24 × 6
4	卷积层: 3 × 3 卷积核,16 通道	24 × 24 × 6 /22 × 22 × 16
5	非线性激活函数 *ReLU*	
6	池化层: 步长为 2 × 2	22 × 22 × 16 /11 × 11 × 16
7	卷积层: 3 × 3 卷积核,32 通道	11 × 11 × 16 /9 × 9 × 32
8	线性激活函数 *ReLU*	
9	池化层: 步长为 2 × 2	9 × 9 × 32 /5 × 5 × 32
10	全连接层	5 × 5 × 32 /120
11	Softmax 层(输出层)	120 /6

3 实验结果与分析

本文使用 2. 3. 2 节所设计的小样本集和 CASIA - HWDB1. 1 数据集作为测试集分别进行了 4 组实验,其中每组实验均包含一组对照实验(简单复制) ,以及一组不作数据增强处理的 0 号实验。分别使用传统数据增强方法、DCGAN 方法和 X - DCGAN 方法生成的增强训练集数据,通过自定义的简单 CNN 分类器进行识别训练和测试,然后比对分析识别效果。实验中使用两种不同分布的测试集,除 0 号实验外(训练集规模为 444) ,其他组训练集规模均为 44 400。除此之外,传统数据增强方法产生的预扩展集输入简单 CNN 训练,得到的模型对 DCGAN 方法生成的数据进行识别,将结果作为扩展集标签参考。实验环境: CPU 为 Intel Core i7 - 7700HQ,内存为 8 GB,GPU 采用 NVIDIA GeForce GTX 1050Ti,操作系统为 Windows10 64 位,实现平台为 python 环境下 Tensorflow - gpu 框架,其中包括 CUDA10. 0 和 CUDNN10 加速包支持。评价指标: 每组识别进行 i 轮训练,获取每轮的测试识别率 $A(i)$,将其最大值 A_{max} 和平均值 A_{ave} 作为评价数据增强算法的性能。公式定义如下。

$$A_{max} = \max(A(i)) \tag{2}$$

$$A_{ave} = \frac{\sum A(i)}{i} \tag{3}$$

表 4 记录了 9 组实验的识别结果。0、1 组的数据显示,简单复制虽能扩充样本数量,但不能增加样本多样性,无法提高识别率。2、4 组的数据显示,对

样本进行 100 倍的数据增强后,使用增强数据进行训练,模型识别率提升了 5% 以上,充分说明传统数据增强方法与 X－DCGAN 方法对小规模手写体汉字集进行了有效扩充与增强,其中,X－DCGAN 方法分别通过预扩展模块和 DCGAN 模块对数据集依次进行了 10 倍的扩充。

表 4　4 种数据增强方法的分类测试识别率

Table 4　Average accuracy of the four data augmentation methods in a classification test

编号	测试集	数据增强	平均识别率/%	最高识别率/%
0	小样本集	无	90.50	94.17
1	小样本集	简单复制	90.42	94.17
2	小样本集	传统方法	97.33	99.16
3	小样本集	DCGAN	85.50	90.83
4	小样本集	X－DCGAN	95.17	98.33
5	HWDB1.1	简单复制	69.50	72.70
6	HWDB1.1	传统方法	70.92	74.65
7	HWDB1.1	DCGAN	59.28	64.62
8	HWDB1.1	X－DCGAN	66.41	70.47

对比 3 种数据增强方法的性能,在小样本测试集下可以看到传统数据增强方法扩充的数据集具有最高识别率,结合式的 X－DCGAN 方法虽然不是最高,但仍然具有较高质量的数据增强效果,完全可以满足一定的识别需求。未改进的 DCGAN 方法效果最差,一方面说明该方法对于小规模数据集的不适应,另一方面也突出了结合式方法的有效性。此外在 HWDB1.1 测试集上 3 种方法的识别率均有大幅下降,说明它们对分布外的数据适应性较差,都有与原始训练样本存在强关联性的问题。在 DCGAN 模块生成样本的过程中观察到,部分图像之间无法从视觉上直观发现不同,并且不同类别的字符之间生成了一定的相似特征,虽然通过增加预扩展集的方法确实提高了训练样本之间的差异性,但效果有限。这种情况提供了另一个值得继续研究的方向。

图 5 展示了采用 DCGAN 方法改进前后,在不同测试集下平均识别率随训练次数的变化。可以明显观察到在两个测试集上采用单一 DCGAN 方法进行数据增强的测试分类识别率均相对较低,使用通过结合方法 X－DCGAN 生成的数据进行分类测试的识别率均有明显提升,尤其在原始样本测试集上的平均识别率提升了 9.67%,充分说明结合式方法能够有效提高生成数据的质量与网络稳定性。即便在识别效果一般的 HWDB1.1 测试集上该方法也同样具有良好的表现。

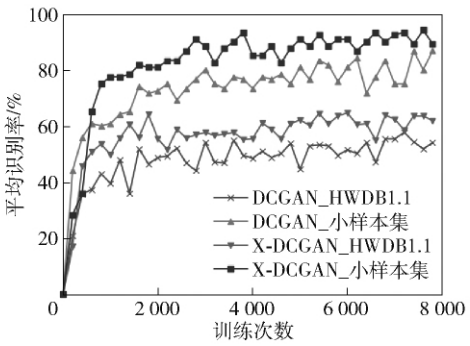

图 5　不同数据集下的平均识别率训练曲线

Fig. 5　Plots of average accuracy of classification for different data sets

4　结论

本文提出结合传统数据增强技术的深度卷积对抗神经网络方法,在进行手写体汉字数据生成时加入预扩展数据,解决了生成式方法在小规模数据集上出现的因样本规模不足导致训练不充分、生成数据重复的问题。对使用结合式方法 X－DCGAN 生成的增强数据集进行分类,结果显示采用该方法不仅能够获得比单一方法更好的分类效果,同时使得生成式方法在小规模手写体汉字数据集下具有很好的生成效果,为扩大数据集规模提供了强有力支持。此外,本文所提的结合方式充分发挥了传统方法和生成式方法各自的优势,既保有快速大量生成样本的能力,同时将图像平均识别率较 DCGAN 方法提升了 9.67%。未来将针对 DCGAN 生成效率和生成图片质量开展进一步研究。

参考文献:

[1] 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. Nevada, 2012: 1097－1105.

[2] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, 2015: 1－9.

[3] HE K M, ZHANG X Y, REN S Q, et al. Deep residual

learning for image recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770 – 778.

[4] WANG S, CHEN L, XU L, et al. Deep knowledge training and heterogeneous CNN for handwritten Chinese text recognition [C] // 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) . Shenzhen, 2016: 84 – 89.

[5] CHEN L, WANG S, FAN W, et al. Beyond human recognition: a CNN-based framework for handwritten character recognition [C] // 2015 3rd IAPR Asian Conference on Pattern Recognition. Kuala Lumpur, 2016: 695 – 699.

[6] LI Z Y, TENG N J, JIN M, et al. Building efficient CNN architecture for offline handwritten Chinese character recognition [J]. International Journal on Document Analysis and Recognition, 2018, 21(4) : 233 – 240.

[7] ZHANG H G, GUO J, CHEN G, et al. HCL2000 —a large-scale handwritten Chinese character database for handwritten character recognition [C] // 2009 10th International Conference on Document Analysis and Recognition. Barcelona, 2009: 286 – 290.

[8] SU T H, ZHANG T W, GUAN D J. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text [J]. International Journal on Document Analysis and Recognition, 2007, 10 (1) : 27 – 38.

[9] JIN L W, GAO Y, LIU G, et al. SCUT-COUCH2009— a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation [J]. International Journal on Document Analysis and Recognition, 2011, 14(1) : 53 – 64.

[10] LIU C L, YIN F, WANG D H, et al. CASIA online and offline Chinese handwriting databases [C] // 2011 International Conference on Document Analysis and Recognition. Beijing, 2011: 37 – 41.

[11] ZHU Y Z, XIE Z C, JIN L W, et al. SCUT-EPT: new dataset and benchmark for offline Chinese text recognition in examination paper [J]. IEEE Access, 2019, 7: 370 – 382.

[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] SIMARD P, STEINKRAUS D, PLATT J C. Best practices for convolutional neural networks applied to visual document analysis [C] // Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, 2003: 958 – 963.

[14] OFUSA K, MIYAZAKI T, SUGAYA Y, et al. Glyph-based data augmentation for accurate kanji character recognition [C] // 2017 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, 2017: 597 – 602.

[15] SONG X C, GAO X, DING Y F, et al. A handwritten Chinese characters recognition method based on sample set expansion and CNN [C] // 2016 3rd International Conference on Systems and Informatics. Beijing, 2016: 843 – 849.

[16] HAYASHI T, GYOHTEN K, OHKI H, et al. A study of data augmentation for handwritten character recognition using deep learning [C] // 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) . Niagara Falls, 2018: 552 – 557.

[17] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] // Advances in Neural Information Processing Systems. Montreal, 2014: 2672 – 2680.

[18] CUI X D, GOEL V, KINGSBURY B. Data augmentation for deep neural network acoustic modeling [C] // 2014 IEEE International Conference on Acoustic, Speech and Signal Processing. Florence, 2014: 5619.

[19] 张晓峰, 吴刚. 基于生成对抗网络的数据增强方法 [J]. 计算机系统应用, 2019, 28(10) : 201 – 206.
ZHANG X F, WU G. Data augmentation method based on generative adversarial network [J]. Computer Systems & Applications, 2019, 28(10) : 201 – 206. (in Chinese)

[20] 陈文兵, 管正雄, 陈允杰. 基于生成式对抗神经网络的数据增强方法 [J]. 计算机应用, 2018, 38(11) : 3305 – 3311.
CHEN W B, GUAN Z X, CHEN Y J. Data augmentation method based on generative adversarial network model [J]. Journal of Computer Applications, 2018, 38(11) : 3305 – 3311. (in Chinese)

[21] ALMÁSI A D, WOŹNIAK S, CRISTEA V, et al. Review of advances in neural networks: neural design technology stack [J]. Neurocomputing, 2016, 174: 31 – 41.

[22] BAIRD H S. Document image defect models and their uses [C] // International Conference on Document Analysis and Recognition. Tsukuba, 1993: 62 – 67.

[23] HORSLEY L, PEREZ-LIEBANA D. Building an automatic sprite generator with deep convolutional generative adversarial networks [C] // 2017 IEEE Conference on Computational Intelligence and Games (CIG) . New York, 2017: 134 – 141.

A data augmentation method based on a small-scale handwritten Chinese character dataset

SHEN Qiang[1]　LI Hui[1*]　ZHANG Yan[2]

(1. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029;
2. Beijing Electrical Engineering School, Beijing 100123, China)

Abstract: In order to improve the poor classification performance of a deep convolutional generative adversarial network(DCGAN) when using a small-scale handwritten Chinese character dataset, a generative method X-DCGAN combined with traditional data augmentation methods is proposed in this work. This method provides more diverse training data for the neural network through the pre-enhancement module, addressing the problem of high sample repetition rate and poor learning effect due to overfitting and insufficient training of the network. Tests showed that the sample data generated by this method have been significantly improved in sample diversity when compared with a single method. In addition, the accuracy obtained by the generating data for classification testing improved by 9.67%. X-DCGAN makes full use of the advantages of traditional augmentation methods and generative methods, and thus can effectively solve the problems of expansion and enhancement of a small-scale dataset.

Key words: data augmentation; deep convolutional generative adversarial networks; handwritten Chinese character recognition; image processing

(责任编辑: 吴万玲)