控制与决策
Control and Decision

基于残差注意网络的端到端手写文本识别方法

王寅同, 郑豪, 常合友, 李朔

引用本文:
王寅同,郑豪,常合友,李朔. 基于残差注意网络的端到端手写文本识别方法[J]. 控制与决策, 2023, 38(7): 1825−1834.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2030

基于残差注意网络的端到端手写文本识别方法

王寅同[1,2†], 郑 豪[1], 常合友[1], 李 朔[3]

(1. 南京晓庄学院 信息工程学院，南京 211171；2. 浙江大学 计算机科学与技术学院，杭州 310058；
3. 英国德蒙福特大学 人工智能研究所，莱斯特 LE19BH)

摘 要: 中文手写文本识别是模式识别领域中的研究热点问题之一,其存在字符类别数量多、书写风格差异大和训练数据集标记难等问题. 针对上述问题,提出无切分无循环的残差注意网络结构用于端到端手写文本识别. 首先,以ResNet-26为主体结构,使用深度可分离卷积提取有意义特征,残差注意门控模块提升文本图像中的关键区域的重要性;其次,采用批量双线性插值模型对输入表征进行拉伸-挤压,实现二维文本表征到一维文本行表征的文本行上采样;最后,以连接时序分类作为识别模型的损失函数,实现高层次抽取表征与字符序列标记的对应关系. 在CASIA-HWDB2.x和ICDAR2013两个数据集上进行实验研究,结果表明,所提方法在没有任何字符或文本行的位置信息时能够有效地实现端到端手写文本识别,且优于现有的方法.

关键词: 手写文本识别；深度可分离卷积；残差注意门控；双线性插值；文本行上采样；连接时序分类

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2030

引用格式: 王寅同,郑豪,常合友,等. 基于残差注意网络的端到端手写文本识别方法[J]. 控制与决策, 2023, 38(7): 1825-1834.

An end-to-end handwritten text recognition method using residual attention networks

WANG Yin-tong[1,2†], *ZHENG Hao*[1], *CHANG He-you*[1], *LI Shuo*[3]

(1. School of Information Engineering，Nanjing Xiaozhuang University，Nanjing 211171，China；2. College of Computer Science and Technology，Zhejiang University，Hangzhou 310058，China；3. Institute of Artificial Intelligence，De Montfort University，Leicester LE19BH，United Kingdom)

Abstract: Handwritten Chinese text recognition which involves thousands of character categories, variant writing styles and monotonous data collection process is a long-standing focus in the field of pattern recognition research. In response to these issues, we propose a residual attention network of segmentation-free and recurrent-free for end-to-end handwritten text recognition with ResNet-26 as the main architecture, using depthwise separable convolution to extract the representation features. The residual attention gate block enhances the important of the key areas of input text image. Then, the text-lines up-sampling of batch bilinear interpolation is used to implement the mapping from two dimension text representation to one dimension text line representation. Finally, connectionist temporal classification as the loss function is employed to realize the corresponding relationship between the high-level extraction representation features and the character sequence labels. An experimental study is carried out on two datasets of CASIA-HWDB2.x and ICDAR2013, and the results indicate that the method can effectively implement end-to-end handwritten text recognition without any position information of characters or text lines, and superior to the existing research methods.

Keywords: handwritten text recognition；depthwise separable convolution；residual attention gate block；bilinear interpolation；text-lines up-sampling；connectionist temporal classification

0 引 言

手写文本识别作为模式识别的一个重要研究领域,自其诞生以来,科研人员从未停止对其进行研究[1-3]. 对于手写文本识别的研究主要归因于以下3个重要方面: 1)作为语音的一种视觉编码形式,手写文本普遍存在于人们的日常生活中,广泛用于交流和

收稿日期: 2021-11-21；**录用日期:** 2022-04-08.
基金项目: 国家自然科学基金项目(62177028, 61976118, 61806098)；江苏省自然科学基金项目(BK20180142)；江苏省青蓝工程项目.
责任编委: 陈家伟.
†通讯作者. E-mail: wangyintong@nuaa.edu.cn.

记录人类的各种思想,包括传统的邮件地址识别、日常办公文档处理和珍贵的历史手稿识别等;2)手写文本的巨大可变性和不断增加的视觉表现形式,导致了手写文本识别问题的复杂性,包括成千上万的字符类别、不同书写者的书写风格差异和字符/文本行的粘连等;3)识别模型需要大量的有标记样本数据,如字符/文本行图像中的字符或字符串标记以及对应的位置信息,这使得样本数据的标记工作成本耗费巨大且易出错. 因此,手写文本识别的应用需求和问题的复杂性,使其成为模式识别领域中的热点研究问题之一.

手写文本识别通常由切分和识别两个阶段构成,面临的问题包括粘连或重叠的文本行或字符难以切分,文本切分的错误累积影响最终识别精度以及模型设计过于复杂等. 由此,从最初的单字符识别[4]衍生出当前主流的文本行识别[5-6],后者可以直接处理输入行图像,无需任何准确的字符或单词切分. 多字符文本行识别能够避免字符的切分错误对识别精度的影响,使得文本行识别性能取得了较大的提升. 与单字符识别不同,文本行识别每次输入一个文本行图像,输出多个字符类别,减少了模型的输入输出次数,同时也能够获得更优的识别率.

在传统手写识别研究中,人们在预处理、特征提取和分类器设计等方面做了大量的努力,取得了不错的研究成果[7]. 然而,近5年内这一类的手写识别方法并没有取得重大的研究进展,鲜有突破性成果. 反之,从深度卷积神经网络赢得了ImageNet图像分类挑战[8]以来,基于深度学习的技术已经广泛应用于大多数计算机视觉相关的任务[9-10]. 深度学习将特征提取和识别两者紧密融合,以较少的领域知识获得更高的识别性能. 文本识别是文本图像到多字符序列的映射,尤其文本行是最典型的一维多字符序列[11-12]. 据作者所知,循环网络的长序列关系在文本识别上的利用并不充分,其原因是文本识别中的长距离字符关系并不显著,反而邻近字符关系对识别结果的影响较大,且模型的循环迭代需要消耗大量计算资源,导致了当前手写文本的识别模型设计向无循环的卷积神经网络发展.

在手写文本识别的无切分和无循环这一研究方向上,残差注意网络的端到端手写文本识别方法被提出,该方法不需要在任何字符/文本行切分信息下进行手写文本的特征抽取,由文本行上采样实现数据表示从二维文本到一维文本行的空间映射,通过连接时序分类的损失函数来引导识别系统的模型训练与文本识别. 本文的主要研究工作包括以下3个内容:1)提出非循环的全卷积神经网络结构,该结构能够有效地避免循环迭代操作带来的大延迟问题,同时卷积操作能够更加充分地利用高性能处理器的并行计算能力. 2)残差注意门控模块设计,能够结合残差和注意力两种机制的优点,突出文本图像中的重要区域和过滤背景中的噪声,缓解深度卷积神经网络的梯度爆炸和梯度消失问题. 3)批量双线性插值模型设计,能够引导二维文本表示上的每个字符特征在不丢失信息的情况下映射到一维文本行表示上的文本行上采样,从而实现文本识别由单行多字符识别到多行多字符识别的转换.

1 相关工作

手写文本识别面临着字符类别多、书写风格多样和字符/文本行结构复杂等问题,一直受到科研工作者的广泛关注,并取得了稳步的发展. 本节从单字符识别、单行多字符识别和多行多字符识别3个方面总结手写文本识别的发展历程,以加深对这一研究领域的理解与认识.

单字符识别是对孤立的手写字符图像进行特征提取,并通过分类器来确定字符类别. 典型的单字符识别模型主要包括预处理、特征提取、降维和字符分类4个阶段[7]. 然而,这类方法受限于预设特征的表示能力,导致识别性能遇到瓶颈. 如改进型二次判别函数作为早期成功的单字符识别方法之一,其性能也很难超过卷积神经网络(CNN)方法,后者在识别速率和识别精度上均取得了显著提升[13]. Li等[14]提出了一种匹配神经网络,从书写汉字的人类学习过程中获得启发,构建了手写字符和模板字符之间的关联关系;Li等[15]运用深度卷积生成对抗网络来实现手写汉字字符识别以改进GoogLeNet方法. 总体而言,单字符识别已取得了良好的识别效果,甚至在具有挑战性的手写字符识别任务上,其识别精度超过了人类的辨别能力. 然而,字符切分作为单字符识别中不可避免的处理过程,复杂的手写文本很难准确地完成字符切分,一旦发生字符切分错误将会导致错误累积效应,给后续的特征提取与字符识别带来严重的影响.

单行多字符识别是从文本行图像到字符序列的识别,通常分为过切分[16-17]和无切分[18]. 前者将文本行图像切分成多个连续的图像片段,对每个图像片段进行识别,并结合语言上下文模型来完成文本行识别. 如Wang等[16]从贝叶斯决策角度引入过切分

方法,通过置信度转换将分类器输出转换为后验概率. Wang等[17]提出了使用异构CNN的深层神经网络,能够从分割候选格中获取分层监督信息. 与过切分方法不同,无切分方法不需要对文本行进行显示切分,而是采用编码-解码或特征对齐等技术实现文本行识别. 如Messina等[19]提出多维度长短时记忆神经网络与循环神经网络方法; Xiao等[20]将像素级校正引入卷积与循环神经网络的文本行识别. 这些方法都是基于循环神经网络,需要大量的计算资源,且在模型训练阶段缺乏并行处理能力. 正因如此,文本识别建模正趋向于无循环神经网络结构. 如Gao等[21]提出了多层卷积堆叠的卷积神经网络以实现端到端文本行识别, Peng等[22]提出了一种用于端到端文本行识别的全卷积网络, Wang等[23]在残差注意力神经网络基础上提出文本行识别方法. 尽管单行多字符识别方法已取得较好的识别性能,但是现有的文本行切分方法对复杂文本结构的处理仍然存在错误,对整个系统的识别性能带来不利影响.

多行多字符识别将手写文本图像识别为一个字符序列,不需要任何字符/文本行的位置信息,能够减少训练样本数据标记的工作量和出错率[24-25]. 传统的识别方法由文本行的检测、 切分和识别3个部分组成. 如Moysset等[26]提出使用递归神经网络和加权有限状态传感器的字符识别系统, Wigington等[27]提出结合区域推荐网络的多行手写文本识别方法, Tensmeyer等[28]提出文本行切分与行识别的方法. 这些方法由不同的预训练模块组成,导致整个系统的预期输出很难实现. 端到端的手写文本识别是通过拉伸-挤压的方式将文本图像映射成几行或一整行的特征表示,新的特征表示能够直接采用单行多字符识别方法进行文本识别. Bluche等[29]提出注意网络的端到端英文手写段落识别; Yousef等[24]提出多行英文文本识别的统一框架,能实现由单行文本识别转换成多行文本识别; Wu等[25]提出基于多维长短时记忆网络的文本识别方法,实现端到端手写中文文本识别. 由于基于隐式分割的方法能够克服多模型组合的手写文本识别不足,近年来得到了广泛的研究和应用.

2　端到端手写文本识别

2.1　算法框架

端到端手写文本识别作为一种无切分和无循环的识别方法,能够将文本图像以端到端的形式识别为对应的字符序列. 其中,注意力机制和残差神经网络的结合能够有效提升文本识别模型从文本图像中提取有意义特征的能力,以及缓解深度卷积神经网络的梯度消失或梯度爆炸问题. 文本行上采样来实现二维文本表征到一维文本行表征的映射,该映射为端到端手写文本识别的一个重要过程. 图1给出了卷积神经网络的手写文本识别流程. 其中:虚线框表示整个特征抽取操作,主要由多残差注意门控堆叠来实现;点线框表示文本行上采样操作,主要由残差注意门控和批量双线性插值来实现.

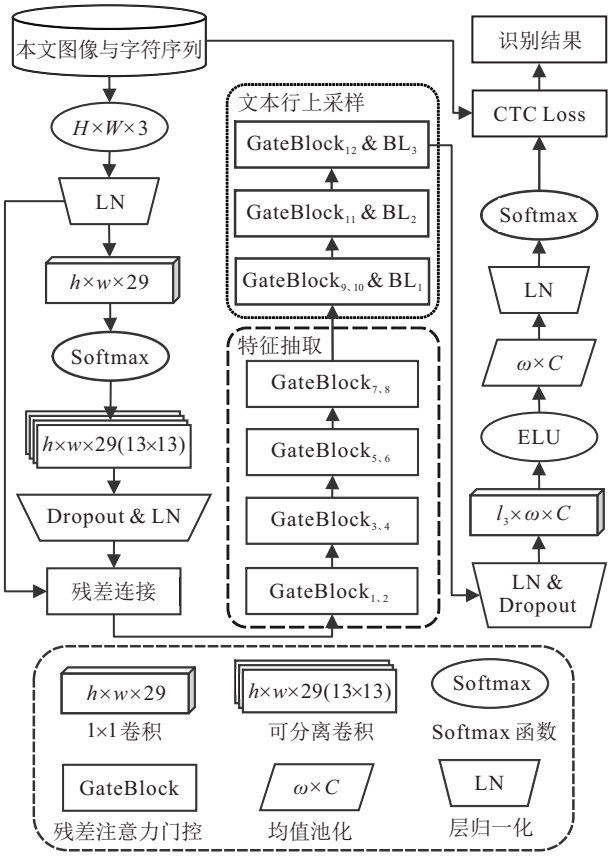

图 1　端到端手写文本识别的流程

从图1可知,本文提出的文本识别方法包括预卷积操作、特征抽取、文本行上采样和文本识别4个阶段,具体内容如下:

1) 预卷积操作. 在保持原图像的宽度和高度的同时增加张量通道数,首先在大小为$h \times w \times 3$的原始图像进行LN层归一化,再执行1×1卷积操作,获得通道数为29的张量;接着引入Softmax归一化以提升数据的收敛速率,并对输入张量中的每个通道分别进行13×13可分离卷积操作;最后,使用Dropout限制部分神经元的输出以在一定程度上实现神经网络的正则化,继续使用LN层归一化,并将输出张量与之前的经过LN层归一化的原始图像进行连接,得到$h \times w \times 32$的输出张量.

2) 特征抽取. 通过堆叠的残差注意门控来实现输入张量的高层次特征抽取,主要由8个残差注意门控和相关池化操作共同实现. 其中8个门控以两两堆叠的形式构成conv1.x ∼ conv4.x四个网络层,并在每层之间增加了相应的池化操作,具体参数如表1所示. 由这4个网络层实现输入张量的特征抽取,张量由$h \times w$大小减小至$h/8 \times w/16$(长度和宽度的最小值设置为8),通道数从32增加至1 024.

表 1 手写整页文本识别的残差注意网络模型参数

	层号	输出大小	操作	次数
特征抽取		$h \times w \times 32$	1×1, Conv2d	
	conv1.x	$\frac{h}{2} \times \frac{w}{2} \times 128$	GateBlock	2
			2×2 max pool, stride 2	
	conv2.x	$\frac{h}{4} \times \frac{w}{4} \times 256$	GateBlock	2
			2×2 max pool, stride 2	
	conv3.x	$\frac{h}{8} \times \frac{w}{8} \times 512$	GateBlock	2
			2×2 max pool, stride 2	
	conv4.x	$\frac{h}{8} \times \frac{w}{16} \times 1\,024$	GateBlock	2
			2×2 max pool, stride 1×2	
文本行上采样	conv5.x	$l_1 \times \frac{w}{32} \times 1\,024$	GateBlock	2
			$l_1 \times \frac{w}{32}$, BL	
	conv6.x	$l_2 \times \frac{w}{64} \times 512$	GateBlock	1
			$l_2 \times \frac{w}{64}$, BL	
	conv7.x	$l_3 \times \frac{w}{128} \times 256$	GateBlock	1
			$l_3 \times \frac{w}{128}$, BL	
文本分类		$l_3 \times \omega \times C$	1×1, Conv2d	
		$l_3 \times 1 \times C$	average pool & LN	
		1	Softmax & CTC	

3) 文本行上采样. 通过残差注意门控引导批量双线性插值实现二维文本表示映射到一维文本行表示的文本行上采样,主要由4个残差注意门控、3个批量双线性插值和相关池化操作共同实现. 由表1可知,文本行上采样主要由cov5.x ∼ cov7.x三个网络层实现. 其中: conv5.x中包括两个残差注意门控和一个批量双线性插值,conv6.x和conv7.x分别包括一个残差注意门控和一个批量双线性插值. 另外,在conv5.x ∼ conv7.x三个网络层上,张量的长度由l_1增加至l_3,宽度由$w/16$减小至$w/128$(长度和宽度的最小值设置为8),通道数由1 024减小至256.

4) 文本识别. 对原始文本图像中提取的表征序列进行识别,其中连接时序分类(CTC)[6]作为文本识别模型的最顶层操作,能够在未切分情况下对两个一维序列的对齐关系计算,实现从表征序列到字符序列的分类,同时也能够利用其空间模型和线性先验知识来引导模型进行二维文本到一维文本行的上采样. 在这一阶段中,对输入张量进行$1 \times 1 \times w$卷积操作获得新的张量,大小为$l_3 \times w \times C$,并在w方向上进行均值池化以获得待识别的表征向量,大小为$l_3 \times 1 \times C$.

2.2 残差注意门控模型

残差注意门控作为手写整页文本识别方法的核心计算单元,既注重输入图像中的有意义特征提取,突出文本图像中的重要区域和过滤背景噪声,又注重深度神经网络训练的收敛性,缓解深度卷积神经网络的梯度爆炸和梯度消失的问题. 设输出张量集$T = \{t_1, t_2, \ldots, t_m\}$. 其中: 第$i$个张量$t_i \in \boldsymbol{R}^{h' \times w' \times 2^k}$, h'、

w'和k是由对应网络层决定的,具体数值见表1. H_c表示一个网络层的输入张量到输出张量的映射函数,即第i网络层的输入张量t_i,其输出张量$t_{i+1} = H_c(t_i)$. 为了实现门控功能,提出改进型残差注意门控映射函数$H_c(t_i)$,其形式化表达式为

$$H_c(t_i) = [M_c(t_i) + 1] \times F_c(t_i) + t_i. \tag{1}$$

其中: t_i表示第i网络层的输入张量, $H_c(t_i)$表示输入张量t_i的c个通道上分别进行映射操作, $M_c(t_i)$表示注意力模块的掩码分支, $F_c(t_i)$表示注意力模块的主干分支.

进一步对式(1)进行恒等变换,得到的形式化表达式为

$$H_c(t_i) = M_c(t_i) \times F_c(t_i) + F_c(t_i) + t_i. \tag{2}$$

对于式(2),令$H'_c(t_i) = M_c(t_i) \times F_c(t_i)$, $H''_c(t_i) = F_c(t_i) + t_i$,得到新的表达式如下:

$$H_c(t_i) = H'_c(t_i) + H''_c(t_i). \tag{3}$$

式(3)由两个部分组成. 其中: $H'_c(t_i)$表示注意力模块,对于第i层网络,从输入张量t_i到输出张量t_{i+1}过程中包含$M_c(t_i)$和$F_c(t_i)$两个分支, $M_c(t_i)$表示注意力模块的掩码分支, $F_c(t_i)$表示注意力模块的主干分支,两者共同作用在特征抽取过程中,强化相关特征信息和抑制不相关特征信息. $H''_c(t_i)$表示残差网络,也同样由两个分支组成,分别为$F_c(t_i)$和t_i, $F_c(t_i)$表示经映射函数(卷积操作)得到的输出张量, t_i表示不经任何处理直接传到下一层网络的输出张量,并作为下一层网络输入张量的一部分.

残差注意门控中需要考虑门控结构内部的维度问题,从式(1)的右侧运算可知,首先将$M_c(t_i)$与1进行相加,接着与卷积函数$F_c(t_i)$进行相乘,最后再与输入张量t_i相加,得到第i层网络的输出张量t_{i+1}. 为此,引入前向变换映射P_1和后向变换映射P_2. 其中: P_1将大小$\boldsymbol{R}^{h\times w\times C}$的张量映射至$\boldsymbol{R}^{h'\times w'\times C'}$,实现张量适应于残差注意门控单元的卷积操作; P_2将大小$\boldsymbol{R}^{h'\times w'\times C'}$的张量映射至$\boldsymbol{R}^{h\times w\times C}$,实现不同网络层之间的张量传输. 将$P_1$和$P_2$代入式(1)中,得到

$$\begin{cases} t'_i = P_1(t_i), \\ t_{i+1} = P_2([M_c(t'_i) + 1] \times F_c(t'_i)) + t_i. \end{cases} \tag{4}$$

其中: t_i表示输入张量, t'_i表示将输入张量t_i进行P_1映射操作, t_{i+1}表示第i网络层的输出张量. 将前向变换映射P_1应用在输入张量t_i上,既能够保留残差连接的优化优势,又能够根据不同的计算资源进行上采样或下采样操作,使得残差注意门控抽取出有意义的特征.

P_1和P_2是通过深度可分离卷积操作来实现,并且引用指数线性单元(ELU)[30]作为激活函数. 令$C' = \alpha C, h' = h, w' = w$,这也意味着,仅在通道维度上进行上采样或下采样操作,而张量的长度和宽度保持不变. α表示张量维度的扩展因子,是以2为底的指数函数,即$\alpha = 2^{k'}, k' \in [-3, -2, -1, 0, 1]$. 当$k' = 0$时,张量的通道数在整个残差注意门控单元中保持不变; 当$k' < 0$时,张量通道数减小,门控的执行速率更高,内存需求更少; 当$k' > 0$时,张量通道数增大,门控的可用信息更多,特征表示能力更好.

2.3　文本行上采样模型

文本行上采样作为解决整页文本识别中在一个垂直方向出现多个字符识别的问题,将输入二维文本表示映射至一维文本行表示,鼓励输入张量的每一行被映射到输出垂直维度的不同部分,且确保后者的空间应足够大以容纳前者映射到后者后能显著区分. 对于包含L个字符的原始文本图像,通过表示提取获得的二维空间的真实序列长度为L,则对应的一维上采样空间的序列长度为L',且$L' \leqslant L$. 换言之,一维上采样空间应该足够大,以使得原始二维空间上的序列和一维上采样空间上的序列能够呈现一对多的关系,并且保持原始序列之间的相邻关系. 为此,在文本行上采样操作中引入批量双线性插值方法,通过长度为l_1、l_2和l_3的张量渐近式地实现二维文本行表征至一维文本行表征的映射.

双线性插值[31]作为一种最优插值方法,对于任意位置的数值估计是由该位置最邻近的4个位置的数值决定,而批量双线性插值是在传统双线性插值基础上提出的一种单次多点插值计算方法,实现更有效的双线性插值操作. 设原始值I,其横坐标为$X = x_1, x_2, \ldots, x_m$,纵坐标为$Y = y_1, y_2, \ldots, y_n$. 新的插入值$I'$,其横坐标为$X' = x'_1, x'_2, \ldots, x'_{m-1}$,纵坐标为$Y' = y'_1, y'_2, \ldots, y'_{n-1}$. 在任意点$(x'_i, y'_j)$处找到插值,首先需要定位该点的4个邻近网格点,即$(x'_i, y'_j)$一定处于横坐标$x_i$、$x_{i+1}$之间和纵坐标$y_i$、$y_{i+1}$之间,其中$0 \leqslant i < m, 0 \leqslant j < n$. 因此,对于任意位置的插值向量$I'$的4个邻近点向量$X_1$、$X_2$、$Y_1$和$Y_2$,对应的值向量为$I_{11}$、$I_{12}$、$I_{21}$和$I_{22}$.

对于4个点向量,采用最佳插值是双线性表示形式:

$$I' = A + BX' + Y'C + DY'X'. \tag{5}$$

系数A、B、C、D由周围的向量值确定,并得到如下关系式:

$$\begin{cases} I_{11} = A + BX_1 + Y_1C + DY_1BX_1, \\ I_{12} = A + BX_1 + Y_2C + DY_2BX_1, \\ I_{21} = A + BX_2 + Y_1C + DY_1BX_2, \\ I_{22} = A + BX_2 + Y_2C + DY_2BX_2. \end{cases} \tag{6}$$

对上面4个等式两两相减,得到

$$\begin{cases} I_{11} - I_{12} = Y_1C + DY_1X_1 - Y_2C - DY_2X_1, \\ I_{21} - I_{22} = Y_1C + DY_1X_2 - Y_2C - DY_2X_2. \end{cases} \tag{7}$$

对上面两个等式相减,得到仅含一个未知系数D的等式,即

$$\begin{aligned} &I_{11} - I_{12} - I_{21} + I_{22} = \\ &DY_1X_1 - DY_2X_1 - DY_1X_2 + DY_2X_2 = \\ &D(Y_1 - Y_2)(X_1 - X_2). \end{aligned} \tag{8}$$

根据矩阵的可逆性质,以及(Y_1-Y_2)和(X_1-X_2)是非零向量,确定$(Y_1 - Y_2)(X_1 - X_2)$存在逆矩阵为$[(Y_1 - Y_2)(X_1 - X_2)]^{-1}$. 对式(8)进行变换操作,得到系数$D$的运算表示式,即

$$D = \frac{I_{11} - I_{12} - I_{21} + I_{22}}{(Y_1 - Y_2)(X_1 - X_2)}. \tag{9}$$

同理,得到系数A、B和C的表示式分别为

$$\begin{cases} A = I_{11} - BX_1 - CY_1 - DY_1X_1, \\ B = (I_{11} - I_{21})(X_1 - X_2)^{-1} - DY_1, \\ C = (I_{11} - I_{12})(Y_1 - Y_2)^{-1} - DX_1. \end{cases} \tag{10}$$

通过给定4个系数向量A、B、C和D的值,新张量I在其4个网格点内的所有点的值可以从式(6)中获得. 重复这个过程,可以得到任意大小的插值向量.

2.4 算法实现

表1给出了端到端手写文本识别的残差注意网络模型的详细参数. 预处理阶段由1×1卷积、可分离卷积和层归一化来实现3通道的原始图像到32通道的张量映射;特征抽取阶段由残差注意门控堆叠构成的conv1.x ~ conv4.x四个网络层来实现;文本行上采样由残差注意门控和批量双线性插值的conv5.x ~ conv7.x三个网络层来实现;文本识别阶段,进行1×1卷积操作,获得长度为$l_3 \times w \times C$向量. 其中:l_3表示张量的长度,$w = 3$表示张量的最终设定宽度,C表示字符类别数. 紧接着,进行宽度w方向的均值池化,获得$l_3 \times 1 \times C$的输出张量. 最后,由Softmax和CTC函数获得表征序列的字符类别,实现模型训练或文本识别.

算法1给出了第i个残差注意门控算法实现,4个输入值分别为输入张量t_i、通道数c、卷积核大小k、扩展因子α. 在算法的实现中,torch包中定义了张量上的相关数学运算以及3个自定义函数:张量复制函数nGates$(\cdot)$、前向深度可分离卷积函数forwardConv$(\cdot)$和后向深度可分离卷积函数blackwardConv$(\cdot)$. 最后,t_{i+1}作为输出值返回.

算法1 第i个残差注意门控算法.

输入: t_i, c, k, α;

输出: t_{i+1}.

1) Def gateBlock(t_i, c, k, α):

2) $t_i' = \text{forwardConv}(t_i, c, k, \alpha)$

3) $t_i'' = \text{torch.ELU}(t_i')$

4) $x_0, x_1 = \text{nGates}(t_i'', c, k, \alpha)$

5) $x_0 = \text{torch.tanh\_}(x_0)$

6) $x_1 = \text{torch.sigmoid}(x_1)$

7) $x = (x_0 + 1)x_1$

8) $z_i = \text{backwardConv}(x, c, k, \alpha)$

9) $z_i' = \text{torch.ELU}(z_i)$

10) $t_{i+1} = z_i' + t_i$

11) return t_{i+1}

算法2给出了端到端手写文本识别算法实现. 除了算法1中的参数外,还包括原始图像I、最大池化操作步长s、文本行上采样长度数组l、张量的最终设定宽度w、字符类别数C和识别结果out. 两个自定义函数包括:残差注意门控函数gateBlock$(\cdot)$、批量双线性插值函数BL$(\cdot)$、预处理卷积函数preConv$(\cdot)$和后处理卷积函数postConv$(\cdot)$. 其中后处理卷积函数主要实现卷积、维度上的平均池、层归一化和Softmax. 然后,使用CTC函数计算原始文本图像抽取得到的表征向量与训练中的真实字符序列的损失值或预测中的文本图像所对应的预测字符序列.

算法2 端到端手写文本识别算法.

输入: $I, c, k, \alpha, s, l, \omega, C$;

输出: out.

1) $t_0 = \text{preConv}(I)$

2) for $i \leftarrow 0$ to 3

3) $t_i' = \text{gateBlock}(t_i, c, k, \alpha)$

4) $t_i'' = \text{gateBlock}(t_i', c, k, \alpha)$

5) $t_{i+1} = \text{maxPool}(t_i'', s)$

6) $t_5 = \text{gateBlock}(t_4, c, k, \alpha)$

7) for $j \leftarrow 0$ to 2

8) $\quad t'_j = \mathrm{gateBlock}(t_{5+j}, c, k, \alpha)$

9) $\quad t_{6+j} = \mathrm{BL}(t'_j, l_j)$

10) $t_9 = \mathrm{postConv}(t_8, k, \omega, C)$

11) $t_{10} = \mathrm{torch.mean}(t_9)$

12) $t_{11} = \mathrm{torch.layer\_norm}(t_{10})$

13) $t_{12} = \mathrm{torch.sigmoid}(t_{11})$

14) $\mathrm{out} = \mathrm{CTC}(t_{12})$

15) return out

3　实验结果与分析

3.1　实验数据集和参数设置

本文使用两个公开数据集与主流的文本识别方法进行实验对比,深入分析所提整页文本识别方法的性能. 其中,CASIA-HWDB2.x[32]由1 019名书写者完成,每位书写者完成5份手稿,整个数据集含有手稿的总数量为5 091 (已排除4份丢失手稿),字符类别的数量为2 703. 数据集分为训练集和测试集两个部分,其中训练集中的手稿数量为4 076,行数为41 781,字符数为1 081 508;测试集中的手稿数量为1 015,行数为10 449,字符数为267 906. ICDAR2013[33]数据集来源于2013年中文手写识别大赛,该数据集包含文本数为300,字符数为91 519,字符类别数为2 703,且全部包含在CASIA-HWDB2.x的字符类别.

本文在Ubuntu16.04、NVIDIA GeForce GTX 2080Ti实验条件下,以ResNet26为主体结构,使用CUDA加速,利用TensorFlow构建基础网络. 在训练模型之前对数据进行预处理,可以扩增数据集,提高网络的泛化能力. 移除原数据集中的文本图像的上部和下部空白区域,将所有文本图像大小调整到$2\,100 \times 2\,400$,以满足网络对输入图像的要求. 模型的初始化学习率设置为1×10^2,并且学习率随着epoch次数的增加呈现下降趋势,两者之间是指数衰减关系,且在epoch次数为1×10^6时,学习率为1×10^3;单次epoch中允许的最小样本数为2,网络层之间的最大池化操作如表1所示. 模型训练的停止条件是最大迭代次数为1×10^6,或损失函数值连续50次迭代不减少.

Levenstein编辑距离[34]是常用于字符级别评估手写识别模型性能的一种度量方式,可以通过标签序列的长度对其进行归一化,也称为字符错误率. 参照现有方法[6,16]的实验评估,以准确率(AR)和正确率(CR)来评估实验比较方法的性能,其形式化表示式如下:

$$\mathrm{AR} = (N_t - D_e - I_e - S_e)/N_t, \tag{11}$$

$$\mathrm{CR} = (N_t - D_e - S_e)/N_t. \tag{12}$$

其中: N_t表示原始文本图像对应的字符序列的长度, S_e表示替换错误字符的数量, D_e表示删除错误字符的数量, I_e表示插入错误字符的数量.

3.2　实验结果对比分析

3.2.1　扩展因子的对比分析

扩展因子在残差注意门控计算中的通道数控制上起着至关重要的作用,将原始输入张量进行上采样或下采样到高维或低维表示空间,并将轻量级深度卷积作用于新张量上. 本实验中的文本行上采样的长度关系设定为模式1和模式2. 对于模式1而言,相邻层的上采样长度关系是后者为前者的两倍,并满足大于或等于最接近100的整数倍. 如$l_3 = 2\,400$,则前两层的上采样长度分别为$l_2 = \lceil l_3/2 \rceil = 1\,200$和$l_1 = \lceil l_2/2 \rceil = 600$. 类似地,模式2中相邻层的上采样长度关系是后者为前者的3/2倍. 表2给出了CASIA-HWDB2.x数据集上文本识别方法在不同扩展因子的准确率(AR)、单次epoch时间(time)和训练模型大小(size). 其中,文本行上采样的长度关系为模式1且$l_3 = 2\,400$,列代表了不同的扩展因子值. 从表2可知,准确率、单次epoch时间和训练后的模型大小随着扩展因子值的不同而变化. 其中:准确率增长较为平缓并趋于一定值,最小值和最大值分别为78.47 %和91.08 %,变化幅度为16.07 %. 然而,单次epoch时间从1 390 s增加到3 795 s,训练模型大小从26.50 MB增加到406.70 MB,这两者的增长速率都远大于准确率的增长速率.

表2　CASIA-HWDB2.x数据集上的整页文本识别模型的扩展因子分析

参数	扩展因子α				
	1/8	1/4	1/2	1	2
AR/%	78.47	86.65	90.53	90.96	91.08
time/s	1 390	1 494	1 670	2 196	3 795
size/MB	26.50	37.80	66.50	147.90	406.70

图2给出了CASIA-HWDB2.x数据集的整页文本识别模型性能趋势. 其中:最终文本行上采样长度为2 400,实线表示文本行上采样长度模式1的参数值,虚线表示文本行上采样长度模式2的参数值.

从图2可知,增大扩展因子值可以提高准确率,但也增加了模型训练中计算资源需求,并且后者呈现

快速增长趋势. 再有,上采样长度模式2的准确率、单次epoch时间和训练模型大小在一定程度上大于上采样长度模式1,这表明当最终上采样长度固定时,间隔小的上采样长度模式更有可能获得较高的准确率.

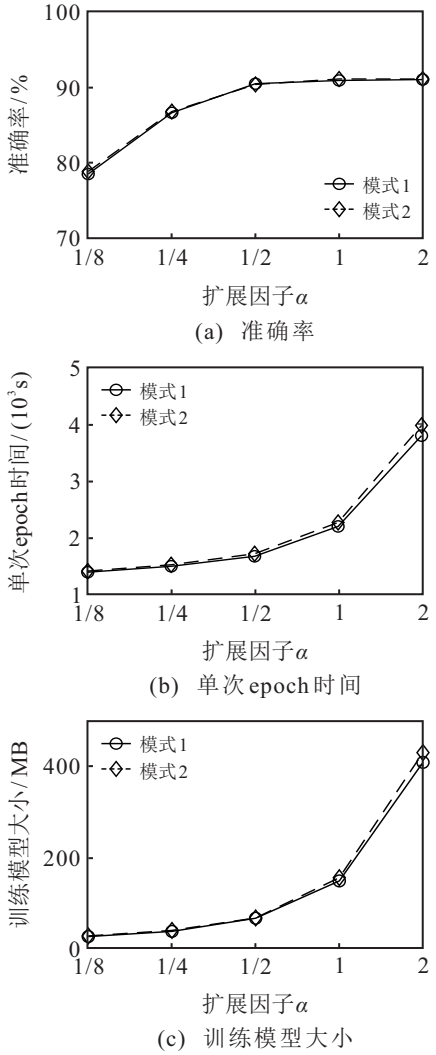

图2 CASIA-HWDB2.x数据集的整页文本识别模型性能趋势

图3给出了CASIA-HWDB2.x数据集的整页文本识别模型损失函数值变化趋势. 其中:文本行上采样的长度关系为模式1,最终文本行上采样长度为2 400,扩展因子值为1/2.

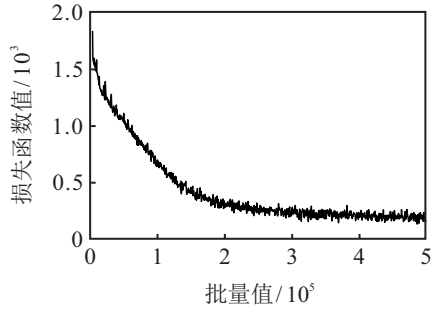

图3 CASIA-HWDB2.x数据集的整页文本识别模型损失函数值变化趋势

从图3可知,最初的损失函数值最大为1 837,随着批量值的增大,损失函数值呈现快速下降趋势;当训练错误率降低后,损失函数值的下降速率变小且趋于平缓.

3.2.2 识别精度及分析

针对本文方法的有效性讨论,参与比较的方法包括Bluche等[29]和Wu等[25]提出的两种基于MLSTM的方法、文献[12]的ResNet-26以及文献[24]的DrigamiNet. 基于MLSTM的方法利用了手写文本图像的上下文信息,但面临递归操作带来的时间成本和空间占用问题. 相反,ResNet-26[12]、OrigamiNet[24]和本文方法都基于CNN,注意机制的引入能够使得模型在特征抽取过程中更加偏向于辨识能力强的特征,以及2D文本到1D文本行的转换,并由连接时序分类实现文本识别.

表3给出了不同方法的文本识别结果,其中"–"表示对应方法的识别准确率缺失. 在CASIA-HWDB2.x数据集上,ResNet-26和OrigamiNet的识别准确率分别为79.25 %和81.72 %,而本文方法实现的最优准确率为90.53%,较前两者分别提高14.23 %和10.78 %. 在ICDAR2013数据集上,Bluche等和Wu等提出的方法获得的准确率分别为68.32 %和80.09 %,后者的准确率较前者提高了11.77 %;其他3种基于CNN方法的准确率分别为68.50 %、71.22 %和81.40 %. 对于所有参与比较的整页文本识别方法,本文方法在CASIA-HWDB2.x和ICDAR2013数据集上均获得了最高准确率,且明显高于前3种方法. 其原因包括模型结构的优化有助于识别准确率的提高,以及前3种方法最初应用于英文整页文本识别,而中文整页文本的类别数多且单个字符的结构更复杂.

表3 CASIA-HWDB2.x和ICDAR2013数据集上的手写整页文本识别精度 %

方法	CASIA-HWDB2.x		ICDAR2013	
	AR	CR	AR	CR
文献[29]方法	–	–	68.32	–
文献[12]方法	79.25	80.61	68.50	69.80
文献[24]方法	81.72	83.54	71.22	72.98
文献[25]方法	–	–	80.09	–
本文方法	**90.53**	**92.60**	**81.40**	**82.85**

4 结 论

针对手写中文文本识别中存在的问题,本文提出了残差注意网络的端到端手写文本识别方法,作为

一种无切分和非循环的识别模型,能够在没有任何字符/文本行的位置信息下实现手写文本识别. 残差注意门控能够增强文本图像中笔迹像素的相关特征信息和抑制背景噪声像素的不相关特征信息;文本行上采样能够实现二维文本表示到一维文本行表示的映射;最终由连接时序分类的损失函数来引导识别系统的模型训练与文本识别. 在所有参与比较的方法中,本文提出的方法在两个实验数据集上均获得最优的识别精度. 在未来的工作中,将重点关注手写文本识别的无切分无循环网络结构的轻量化设计,并引入对抗生成网络来获得更多高质量的扩增数据以实现识别模型的充分训练.

参考文献(References)

[1] Tan Y F, Connie T, Goh M K O, et al. A pipeline approach to context-aware handwritten text recognition[J]. Applied Sciences, 2022, 12(4): 1870.

[2] Wang Z R, Du J, Wang J M. Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition[J]. Pattern Recognition, 2020, 100: 107102.

[3] 金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述[J]. 自动化学报, 2016, 42(8): 1125-1141.
(Jin L W, Zhong Z Y, Yang Z, et al. Applications of deep learning for handwritten Chinese character recognition: A review[J]. Acta Automatica Sinica, 2016, 42(8): 1125-1141.)

[4] 王恺, 李成学, 王庆人, 等. 异态汉字识别方法研究[J]. 软件学报, 2014, 25(10): 2266-2281.
(Wang K, Li C X, Wang Q R, et al. Research on abnormal Chinese character recognition[J]. Journal of Software, 2014, 25(10): 2266-2281.)

[5] Yousef M, Hussain K F, Mohammed U S. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks[J]. Pattern Recognition, 2020, 108: 107482.

[6] Xie C Y, Lai S X, Liao Q Y, et al. High performance offline handwritten Chinese text recognition with a new data preprocessing and augmentation pipeline[C]. International Workshop on Document Analysis Systems. Wuhan, 2020: 45-59.

[7] Wei X H, Lu S J, Lu Y. Compact MQDF classifiers using sparse coding for handwritten Chinese character recognition[J]. Pattern Recognition, 2018, 76: 679-690.

[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[9] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报, 2019, 42(3): 453-482.
(Zhang S, Gong Y H, Wang J J. The development of deep convolution neural network and its applications on computer vision[J]. Chinese Journal of Computers, 2019, 42(3): 453-482.)

[10] Liu B, Xu X C, Zhang Y. Offline handwritten Chinese text recognition with convolutional neural networks[J/OL]. 2020, arXiv: 2006.15619.

[11] Wang Z R, Du J. Joint architecture and knowledge distillation in CNN for Chinese text recognition[J]. Pattern Recognition, 2021, 111: 107722.

[12] 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.

[13] 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: IJDAR, 2018, 21(4): 233-240.

[14] Li Z Y, Wu Q, Xiao Y, et al. Deep matching network for handwritten Chinese character recognition[J]. Pattern Recognition, 2020, 107: 107471.

[15] Li J W, Song G, Zhang M H. Occluded offline handwritten Chinese character recognition using deep convolutional generative adversarial network and improved GoogLeNet[J]. Neural Computing and Applications, 2020, 32(9): 4805-4819.

[16] Wang Q F, Yin F, Liu C L. Handwritten Chinese text recognition by integrating multiple contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1469-1481.

[17] Wang S, Chen L, Xu L, et al. Deep knowledge training and heterogeneous CNN for handwritten Chinese text recognition[C]. The 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). Shenzhen, 2016: 84-89.

[18] 姚红革, 董泽浩, 喻钧, 等. 深度EM胶囊网络全重叠手写数字识别与分离[J]. 自动化学报, DOI: 10.16383/j.aas.c190849.
(Yao H G, Dong Z H, Yu J, et al. Fully overlapped handwritten number recognition and separation based on deep EM capsule network[J]. IEEE/CAA Journal of Automatica Sinica, DOI: 10.16383/j.aas.c190849.)

[19] Messina R, Louradour J. Segmentation-free handwritten Chinese text recognition with LSTM-RNN[C]. The 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, 2015: 171-175.

[20] Xiao S Y, Peng L R, Yan R J, et al. Deep network with pixel-level rectification and robust training for

handwriting recognition[J]. SN Computer Science, 2020, 1(3): 1-13.

[21] Gao Y Z, Chen Y Y, Wang J Q, et al. Reading scene text with fully convolutional sequence modeling[J]. Neurocomputing, 2019, 339: 161-170.

[22] Peng D Z, Jin L W, Wu Y Q, et al. A fast and accurate fully convolutional network for end-to-end handwritten Chinese text segmentation and recognition[C]. International Conference on Document Analysis and Recognition (ICDAR). Sydney, 2019: 25-30.

[23] Wang Y T, Yang Y J, Ding W P, et al. A residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks[J]. IEEE Access, 2021, 9: 132301-132310.

[24] Yousef M, Bishop T E. OrigamiNet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 14698-14707.

[25] Wu Y C, Hu X L. From textline to paragraph: A promising practice for Chinese text recognition[C]. Proceedings of the Future Technologies Conference. Vancouver, 2021: 618-633.

[26] Moysset B, Bluche T, Knibbe M, et al. The A2iA multi-lingual text recognition system at the second maurdor evaluation[C]. The 14th International Conference on Frontiers in Handwriting Recognition. Hersonissos, 2014: 297-302.

[27] Wigington C, Tensmeyer C, Davis B, et al. Start, follow, read: End-to-end full-page handwriting recognition[C]. Proceedings of the European Conference on Computer Vision. Munich, 2017: 367-383.

[28] Tensmeyer C, Wigington C. Training full-page handwritten text recognition models without annotated line breaks[C]. International Conference on Document Analysis and Recognition (ICDAR). Sydney, 2019: 1-8.

[29] Bluche T, Louradour J, Messina R. Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention[C]. The 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, 2017: 1050-1055.

[30] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs)[J/OL]. 2015, arXiv: 1511.07289.

[31] Khosravi M R, Samadi S. BL-ALM: A blind scalable edge-guided reconstruction filter for smart environmental monitoring through green IoMT-UAV networks[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(2): 727-736.

[32] Liu C L, Yin F, Wang D H, et al. CASIA online and offline Chinese handwriting databases[C]. International Conference on Document Analysis and Recognition. Beijing, 2011: 37-41.

[33] Yin F, Wang Q F, Zhang X Y, et al. ICDAR 2013 Chinese handwriting recognition competition[C]. The 12th International Conference on Document Analysis and Recognition. Washington, 2013: 1464-1470.

[34] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals[J]. Soviet Physics Doklady, 1965, 10: 707-710.

作者简介

王寅同 (1987–), 男, 讲师, 博士, 从事维数约简、手写体识别等研究, E-mail: wangyintong@nuaa.edu.cn;

郑豪 (1976–), 男, 教授, 博士, 从事微表情识别、情感交互等研究, E-mail: zhh710@163.com;

常合友 (1991–), 男, 讲师, 博士, 从事稀疏表示、深度学习等研究, E-mail: cv_hychang@126.com;

李朔 (1977–), 男, 讲师, 博士生, 从事大数据处理、教育数据挖掘等研究, E-mail: shuo.li@dmu.ac.uk.