doi: 10.3969/j.issn.0490-6756.2020.02.009

基于深度学习的任意形状场景文字识别

徐富勇, 余 谅, 盛钟松
(四川大学计算机学院, 成都 610065)

摘 要: 场景文字识别的一个具有挑战性的方面是处理具有扭曲或不规则布局的文字. 尤其是侧视文字和曲线文字在自然场景中较为常见,且难以识别. 本文提出了一个带有灵活矫正功能的注意力增强网络,将其用于任意形状场景文字识别. 此网络由基于卷积神经网络的文字矫正网络和基于注意力增强的识别网络两部分组成. 矫正网络自适应地将输入图像中的文字进行矫正,降低识别难度,使基于注意力增强的序列识别网络直接根据矫正后的图像预测字符序列. 整个模型可以进行端到端的训练,训练只需要图像和相应的文字真实标签. 在各种公开数据集上进行了广泛的实验,包括 SVT、ICDAR 2003 和 CUTE80 等数据集,验证了此网络具有优异的性能.

关键词: 深度学习; 场景文字识别; 神经网络; 注意力机制
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0490-6756(2020)02-0255-09

Arbitrary shape scene text recognition based on deep learning

XU Fu-Yong, *YU Liang*, *SHENG Zhong-Song*
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: One of the challenging in scene text recognition is to deal with distortions or irregular layout. Especially, perspective text and curved text are common in natural scenes and are difficult to recognize. In this paper, we propose an attention enhanced network with flexible rectification function for Arbitrary shape scene text recognition. The network consists of a text rectification network and an attention enhanced recognition network. The rectification network adaptively rectifies the text in the input image to reduce the difficulty recognition. The recognition network is an attention enhanced sequence-to-sequence model that predicts a character sequence directly from the rectified image. With end to end training approach, only images and corresponding text labels are required. Extensive experiments have been conducted on a variety of open datasets, including SVT, ICDAR 2003 and CUTE80, and the experimental results shows the proposed network has excellent performance.

Keywords: Deep learning; Scene text recognition; Neural network; Attention mechanism

1 引 言

近年来,由于自然场景文字识别在广泛应用中的重要性,其引起了学术界和工业界的广泛关注. 很多应用都受益于场景文字的丰富语义信息,比如:交通标志的识别[1-2]、产品识别[3-4]、图片搜索和无人驾驶[5]等. 随着场景文字检测方法的发展,场景文字识别也成为当前研究的前沿课题,也是一个

收稿日期: 2019-04-30
基金项目: 国家自然科学基金(61872256)
作者简介: 徐富勇(1995-), 男, 云南昭通人, 硕士生, 研究方向为深度学习. E-mail:1329950774@qq.com
通讯作者: 余谅. E-mail: yuliang@scu.edu.cn

开放性和极具挑战性的研究课题.

目前,规则的文字识别[6]取得了显著的成功.基于卷积神经网络的方法[6]得到了广泛的应用.有很多研究方法将递归神经网络[7-8]和注意机制[9-12]结合到识别模型中,并且还取得了很好的效果.然而,目前大多数的识别模型仍然不稳定,无法处理来自环境的多种干扰.不规则文字的各种形状和扭曲模式对识别造成了更大的困难.如图 1 所示,透视和曲线形状等不规则的场景文字仍然很难识别.

(a) (b) (c)

图 1 规则和不规则场景文字例子
(a) 规则文字;(b)、(c) 不规则文字.
Fig. 1 Examples of regular and irregular scene text
(a) Regular text;(b)~(c) irregular text.

因此,我们提出了一种带有灵活矫正功能的注意力增强网络 FRAEN (Flexible Rectification Attention Enhanced Network),它可以识别自然场景中缩放和拉伸的文字.此网络由灵活矫正网络 FRN (Flexible Rectification Network)和基于注意力增强的网络 AEN (Attention Enhanced Network)的识别网络组成.我们把困难的识别任务分成两部分.首先,FRN 作为一种图像空间转换器,对包含任意形状文字的图像进行矫正.如图 2 所示,经过 FRN 的矫正,倾斜的文字变得更加水平,更容易识别.紧接着,AEN 将矫正后的图像作为输入,直接输出预测的单词.

当前的文字识别网络,那些具有注意力机制的解码器更可能利用经过矫正的图像预测正确的单词.但是 Cheng 等人[9]发现现有的基于注意力的方法会出现注意力偏移的情况.因此,根据他们所提出方法的启发,我们针对自己的模型,提出了注意力增强的方法来改进和训练 AEN.提出了基于相邻注意力权重的双向 GRU(Gated Recurrent Unit)解码器.由于注意力增强的作用,AEN 对于上下文的变化更加鲁棒.简而言之,本文的主要贡献如下:(1) 本文提出的 FRAEN 能够很好地处理和识别不规则的场景文字;(2) 本文提出了一种基于注意力增强的解码器方法,本方法可以解决注意力偏移的问题;(3) 本文提出的方法可以以弱监督的方式进行训练,只需要提供文字标签,这样省去了大量的标注工作.

2 相关工作

近年来,由于神经网络的快速发展[13-15],对规则场景文字的识别能力已经大大提高.文献[11]概述了场景文字检测和识别领域的主要进展.由神经网络提取的模式特征相比于手工制作的特征变得占主导地位,例如 Semi-Markov 条件随机场和生成形状模型.Jaderberg 等人[16]和 Yin 等人[17]使用卷积神经网络 CNNs(Convolutional Neural Networks),提出了无约束识别的各种方法.

随着递归神经网络 RNN(Recurrent Neural-Network)的广泛应用,基于 CNN(Convolutional Neural Network)与 RNN 结合的方法可以更好地学习上下文信息.Shi 等人[18]提出了一个具有 CNN 和 RNN 的端到端可训练网络,称为 CRNN (Convolutional Recurrent Neural Network).此外,注意力机制侧重于信息区域以实现更好的性能.文献[11]提出了一种基于注意力机制的递归网络,用于场景文字识别.Cheng 等人[9]使用聚焦注意网络来纠正注意力机制的变化,实现更准确的注意力位置预测.

与规则场景文字识别工作相比,不规则文字识别更加困难.一种不规则的文字识别方法是自底向上的方法[12],它搜索每个字符的位置然后连接它们.另一种是自顶向下的方法[8]直接从整个输入图像识别文本,而不是检测和识别单个字符.我们提出的 FRAEN 方法采用的是自顶向下的方法.注意力增强方法被用于提高 FRAEN 注意力的准确度.我们使用端到端的方式训练 FRAEN,可以使得文字矫正网络和文字识别网络很好的结合.

3 方 法

FRAEN 包含两部分,FRAEN 的整体架构图如图 2 所示.第一部分是 FRN,在本部分,由于目前提出的矫正网络都仅仅矫正水平方向,在本文中,我们加入一个由基本 CNN 构成的方向标准化网络,将垂直方向的文字转为水平方向文字,统一进行图像矫正,FRN 网络的作用是学习图像每个部分的偏移量,根据学习的偏移量,我们通过双线性插值采样获得矫正后的文字图像;另一部分是

AEN，由带有注意力增强解码器的 CNN-BLSTM (Bi-directional Long Short-Term Memory)-GRU 架构构成. 直接处理和识别矫正后的图像，输出预测结果.

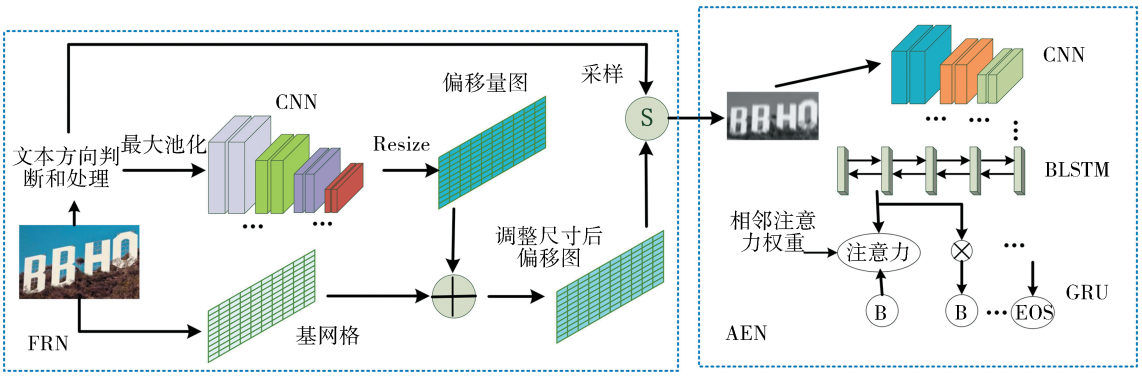

图 2　FRAEN 整体架构

Fig. 2　Overall structure of FRAEN

3.1　FRN

常用的模式矫正方法，如仿射变换网络，其受到一定的几何约束，仅限于旋转，缩放和平移. 然而，一个图像可能有多种变形，尤其自然场景文字的变形更是复杂多变的. 由于识别模型对各种形状的多扰动处理能力不强. 所以，我们考虑对图像进行矫正以降低识别的难度. 如图 2 所示，FRN 架构首先对传入网络的图像进行一个二分类判断，只判断图像中文字是否为垂直方向，并进行旋转处理，将处理后的图像传入由 CNN 构成的矫正网络，进行文字矫正处理. 我们将一个最大池化层放在矫正网络之前，这样可以避免噪声和减少计算量.

表 1　FRN 架构

Tab. 1　Architecture of the FRN

类型	输出形状	配置
输入	1×32×100	–
最大池化	1×16×50	核 2×2，步长 2×2
卷积层	64×16×50	核 3×3，步长 1×1，填充 1×1
最大池化	64×8×25	核 2×2，步长 2×2
卷积层	128×8×25	核 3×3，步长 1×1，填充 1×1
最大池化	128×4×12	核 2×2，步长 2×2
卷积层	64×4×12	核 3×3，步长 1×1，填充 1×1
卷积层	16×4×12	核 3×3，步长 1×1，填充 1×1
卷积层	2×4×11	核 3×3，步长 1×1，填充 1×1
最大池化	2×3×11	–
Resize	2×32×100	–

FRN 架构如表 1 所示. 在我们实现此网络时，除最后一个卷积层外，每个卷积层后面都有一个批处理归一化层和一个 ReLU(Rectified Linear Unit)层. 由表 1 可以看出，首先，FRN 将图像分割为 3×11=33 个部分，预测每个部分的偏移量，输入大小为 32×100，得到的偏移量图包含两个部分，分别代表原图像素在 x 坐标和 y 坐标方向的偏移量；然后，我们使用双线性插值平滑地调整偏移量图，使其与输入图像大小相同都为 32×100.

偏移量图中的每个值表示原图原始位置的偏移量，因此我们先为输入图像生成一个基网格来表示像素的原始位置，该基网格使用 x 和 y 坐标表示输入图像像素的位置. 将每个像素的坐标归一化至[−1,1]. 左上角像素的坐标为(−1,−1)，右下角的坐标为(1,1). 最后，将基网格和得到的偏移量图以如下方式进行求和.

$$\text{offset}'(c,i,j)=\text{offset}(c,i,j)+\text{basic}(c,i,j),c=1,2 \tag{1}$$

公式(1)中，(i,j)代表网格第 i 行和第 j 列的位置. $c=1,2$ 分别代表 x 坐标和 y 坐标. 对于偏移量图而言，对应的是需要对原图 x 和 y 坐标位置的像素进行调整的偏移量. 而对于基网格则是输入图像像素位置的 x 和 y 坐标.

采样前，偏移量图上的 x 坐标和 y 坐标分别归一化到[0,W]和[0,H]. 这里，$H\times W$ 是输入图像的大小. 矫正后的图像 I'的第 i 行和第 j 列的像素值由以下公式得到：

$$I'(i,j)=I(i',j') \tag{2}$$

$$\begin{cases} i'=\text{offset}'(1,i,j) \\ j'=\text{offset}'(2,i,j) \end{cases} \tag{3}$$

其中，I 是输入图像；i'和 j'分别对应于式(1)中 $c=1$ 和 2 的取值. 这里得到的 i'和 j'都是实数，而不是整数，因此，经过矫正的图像 I'，是我们采用双线性插值方法从图像 I 中采样得到. 由于式(2)是

可微的,FRN 可以进行梯度反向传播训练.

如图 3 所示,左边显示未进行矫正处理的不规则文本图像,右边显示的是经过 FRN 矫正处理后的文本图像. 从图 3 可以看出,经过校正的图像中的文本更规则和更具可读性,倾斜和透视的文本经过矫正后变得紧密结合,弯曲文本也变得更规则.

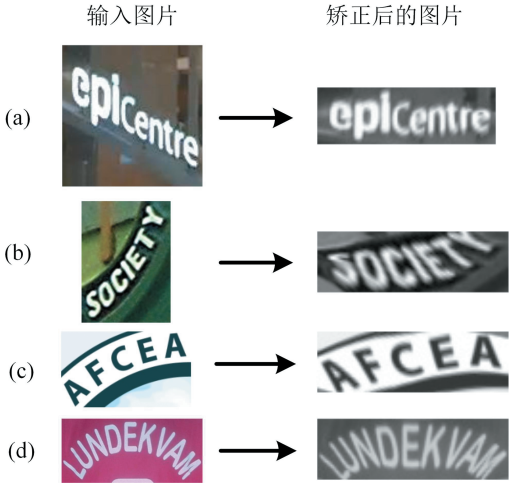

图 3 FRN 矫正不规则文字的结果

Fig. 3 Results of the FRN on irregular text.

3.2 AEN

如图 2 所示,AEN 的主要结构是 CNN-BLSTM-GRU 框架. 编码器部分我们采用的是 CNN-BLSTM 架构. 目前方法的解码器是基于 GRU 直接生成目标序列$(y_1,y_2,\cdots,y_T)$. 解码器生成的最大步数为 T. 解码器在预测到序列结束标记 EOS 时停止处理. 在时间步 t,输出 y_t 如下.

$$y_t=\mathrm{Softmax}(W_{\mathrm{out}}s_t+b_{\mathrm{out}}) \tag{4}$$

式(4)中,s_t 是时间第 t 步隐藏层状态,我们使用 GRU 来更新 s_t,由如下公式计算更新.

$$s_t=GRU(y_{\mathrm{prev}},g_t,s_{t-1}) \tag{5}$$

式(5)中,y_{prev} 代表的是前一个时间段的输出 y_{t-1} 的嵌入向量;g_t 代表注意力权重向量.

$$y_{\mathrm{prev}}=\mathrm{Embedding}(y_{t-1}) \tag{6}$$

$$g_t=\sum_{i=1}^{L}(\alpha_{t,i}h_i) \tag{7}$$

式(7)中,h_i 代表的是序列特征向量;L 是特征图的长度. 而第一项 $\alpha_{t,i}$ 是注意力权重向量,计算如下.

$$\alpha_{t,i}=\frac{\exp(e_{t,i})}{\sum_{j=1}^{L}(\exp(e_{t,j}))} \tag{8}$$

$$e_{t,i}=\mathrm{Tan}\,h(W_ss_{t-1}+W_hh_i+b) \tag{9}$$

在式(4)~式(9)中,W_{out},b_{out},W_s,W_h 和 b 都是可训练的参数. 注意:在训练阶段 y_{prev} 是来自最后一步的真实标记. 然而,在测试阶段使用最后一步的预测输出作为 y_{t-1}. 本文解码器是基于注意力增强的解码器,借鉴文献[5]的思想,本文提出了相邻注意力权重和双向 GRU 解码器方法,在 3.3 节和 3.4 节详细说明. AEN 的架构详细信息见表 2. 编码器部分采用了 45 层的残差网络结构作为卷积神经网络,每个残差单元都由一个 1×1 的卷积层伴随一个 3×3 的卷积层组成. 在第 1 个和第 2 个残差块中,图像被 2×2 步长的卷积层所降采样. 而在最后的三个残差块中,降采样步长变为 2×1,这样能够更好地区分宽度较窄的字母. 卷积神经网络之后是两层的双向 LSTM 网络,其中的每一层都由一对 LSTM 网络组成,LSTM 的隐藏层单元数量均为 256. 解码器是带有注意力机制的 GRU 网络,注意力机制的单元数和隐藏层单元数均为 256.

表 2 AEN 架构

Tab. 2 Architecture of the AEN

类型	网络层	输出形状	配置
编码器	残差块 0	32×100	3×3 conv,步长 1×1
	残差块 1	16×50	$\begin{pmatrix}1\times1\mathrm{conv},32\\3\times3\mathrm{conv},32\end{pmatrix}\times3$,步长 2×2
	残差块 2	8×25	$\begin{pmatrix}1\times1\mathrm{conv},64\\3\times3\mathrm{conv},64\end{pmatrix}\times4$,步长 2×2
	残差块 3	4×26	$\begin{pmatrix}1\times1\mathrm{conv},128\\3\times3\mathrm{conv},128\end{pmatrix}\times6$,步长 2×1
	残差块 4	2×27	$\begin{pmatrix}1\times1\mathrm{conv},256\\3\times3\mathrm{conv},256\end{pmatrix}\times6$,步长 2×1
	残差块 5	1×26	$\begin{pmatrix}1\times1\mathrm{conv},512\\3\times3\mathrm{conv},512\end{pmatrix}\times3$,步长 2×1
	双向 LSTM	26	隐藏层节点数:256
	双向 LSTM	26	隐藏层节点数:256
解码器	注意力 GRU (从左至右)	26	隐藏层节点数:256
	注意力 GRU (从右至左)	26	隐藏层节点数:256

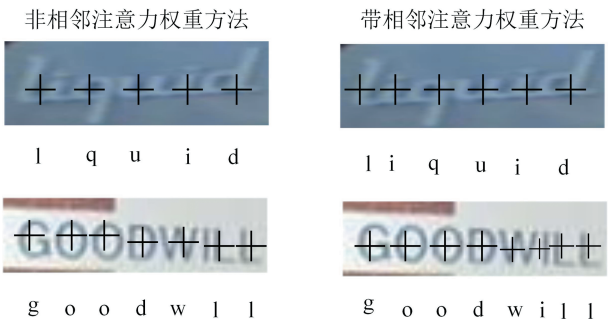

图 4 是否带相邻注意力权重方法训练的比较

Fig. 4 Difference of training with and without adjacent attention weight methods

3.3　相邻注意力权重方法

解码器通过正确注意力的反馈，可以增强选择正确注意力区域的能力. 但是，自然场景图像中存在着各种类型的噪声. 在实际应用中，解码器可能会被欺骗以关注模糊背景区域. 如果解码器生成不正确的注意力区域，选择非对应的特征，这将会导致预测失败. 如图 4 所示，图像包含具有阴影以及复杂背景的文字. 左边的解码器产生了错误的注意力区域，得到了错误的预测结果，遗漏了字母 i.

我们使用了一种称为相邻注意力权重的训练方法，它在训练阶段每一个时间步都获取一对相邻的特征. 通过此方法训练的注意力解码器可以感知相邻的字符. 我们在解码器的每个时间步选择和修改一对注意力. 在时间步 t，$\alpha_{t,k}$ 和 $\alpha_{t,k+1}$ 以如下方式更新.

$$\begin{cases}\alpha'_{t,k}=\beta\alpha_{t,k}+(1-\beta)\alpha_{t,k+1}\\ \alpha'_{t,k+1}=(1-\beta)\alpha_{t,k}+\beta\alpha_{t,k+1}\end{cases}\tag{10}$$

其中，β 是(0,1)间随机生成的小数；k 是$[1,T-1]$间随机生成的整数；T 代表解码器的最大步长.

基于相邻注意力权重方法的解码器，在 $\alpha_{t,k}$ 和 $\alpha_{t,k+1}$ 中都加入了随机性. 这意味着：即使对于相同的图像，在训练阶段的每个时间步长，α_t 的分布都会发生变化. 如式(7)所述，注意力向量 g_t 根据 α_t 的各种分布来获取序列特征向量 h_i，其等同于特征区域在变化. β 和 k 的随机性不仅可以避免过拟合，并且可以增强解码器的鲁棒性. 注意：$\alpha_{t,k}$ 和 $\alpha_{t,k+1}$ 是相邻的. 在不使用相邻注意力权重方法时，序列特征向量 h_k 的误差项是

$$\delta_{h_k}=\delta_{g_t}\alpha_{t,k}\tag{11}$$

上式中，δ_{g_t} 是注意力向量 g_t 的误差项；δ_{h_k} 仅与 $\alpha_{t,k}$ 有关. 但是，使用相邻注意力权重方法，误差项变为

$$\delta_{h_k}=\delta_{g_t}(\beta\alpha_{t,k}+(1-\beta)\alpha_{t,k+1})\tag{12}$$

其中，$\alpha_{t,k+1}$ 与 h_k 相关，如式(8)和式(9)所示，这意味着 δ_{h_k} 受相邻特征决定. 因此，反向传播的梯度能够在更宽范围的相邻区域上动态地优化解码器.

使用上述方法训练的 FRAEN 在每个时间步骤产生更平滑的 α_t. 所以，我们不仅可以提取目标字符的特征，而且还提取了前景和背景上下文的特征. 如图 4 所示，使用此方法能够正确地预测目标字符.

3.4　双向解码器

在我们上述的方法中，使用的序列到序列注意力模型只能捕捉一个方向上的标签相关性. 在实际中，从左到右和从右到左两个方向上的相关性对识别都有利. 例如，一个从左到右工作的解码器可能会因为缺乏上文而难以识别一些单词的首字母，尤其是当该字母为大写‘I’或小写‘l’这样容易混淆的字母时. 相比之下，一个从右到左的解码器则可能更容易识别这些字母，因为它可以根据语言先验知识，由其余字母去推测首字母.

上述的例子表明，工作在相反方向上的两个解码器可能存在互补性. 为了同时利用两个方向上的相关性，我们提出一种双向解码器. 如图 5 所示，双向解码器由两个预测方向相反的解码器构成. 一个从左到右地识别字母序列，另一个从右到左. 从右到左解码器的输出和另一个解码器的输出进行得分比较. 得分较高的标签序列被输出，较低的被丢弃. 这里的得分为解码器每一步的判断得分的累加值. 实际中，我们使用基于贪心算法的解码器，在每一步解码中都选取得分最高的标签作为输出，当输出为 EOS 时停止.

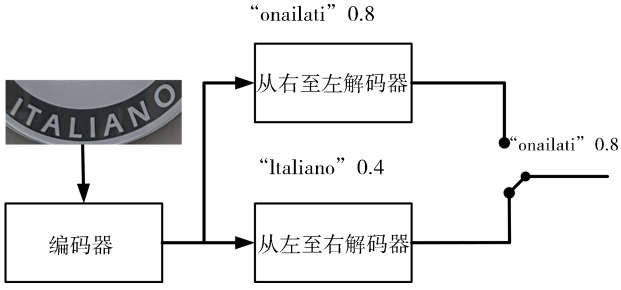

图 5　双向解码器

Fig. 5　Bidirectional decoder

3.5　模型训练

FRAEN 的训练是端到端且是多任务的. 因此，训练的损失函数为

$$L=-\frac{1}{2}\sum_{t=1}^{T}(\log p_{ltr}(y_t\mid I)+\log p_{rtl}(y_t\mid I))\tag{13}$$

其中，$y_1,\cdots,y_t,\cdots,y_T$ 表示标注的字母标签序列. 损失函数为两个解码器(其预测概率分别由 p_{ltr} 和 p_{rtl} 表示)各自损失函数的平均. 等式的右侧只需由图像和标签序列标注计算得到，因此网络的训练只需要图像和对应的标注文字.

模型的所有网络层参数都是随机初始化的. 通过随机梯度下降法进行训练，梯度通过反向传播算法进行计算，我们采用的卷积神经网络和循环神经网络都可以进行反向传播，因此 FRAEN 可以将其接收的误差梯度传递到每一个网络层上，将所有网

络进行端到端训练.

网络训练的优化算法使用 Adadelta,通过 Adadelta 分别计算每个参数上的学习率. 在实际使用中,Adadelta 的收敛速度快.

4 实 验

在本节中,我们在各种基准数据集上进行广泛实验,包括规则和不规则文字数据集. 所有方法的性能都是通过单词级的精度来衡量的. 我们在表 3 中列出了逐步组合本文各方法得到的结果. 可以看出在将所有方法都统一为一个网络结构时,取得了最好的效果.

表 3 FRAEN 的准确率

Tab. 3 Accuracy of the FRAEN

方法	IIIT5k	SVT	IC03	IC13	SVT-P	CUTE80	IC15
无矫正	85.7	82.2	91.4	89.7	71.0	64.6	59.4
有矫正	89.2	87.9	93.7	89.9	74.1	73.2	64.6
单向解码器	89.7	87.4	94.5	90.7	75.5	77.1	68.6
基于相邻注意力权重	91.2	87.9	95.0	91.5	75.9	77.4	70.1
双向解码器	92.2	88.3	95.6	92.4	76.1	79.5	72.7

4.1 数据集

IIIT5K-Words (IIIT5K)[20] 包含用于测试的 3 000张裁剪单词图像. 每张图像都有一个 50 词的词汇表和一个 1 000 词的词汇表. 词汇表由一个正确的单词和其他随机选择的单词组成.

SVT (Street View Text)[19] 采集自 Google Street View,其测试集包含 647 张裁剪后的图片. 许多图片都受到噪声的严重影响,或者分辨率很低. 每个图像都与一个 50 词的词汇表相关联.

ICDAR 2003(IC03)是 ICDAR 2003 竞赛所使用的数据集. 本文只使用其识别数据集. 包含非字母数字和长度小于 3 的文字图片被从数据集中剔除. 过滤后的识别数据集包含 860 张裁剪图片.

ICDAR 2013 (IC13)[21] 的大部分样本都继承自 IC03. 它包含 1015 个裁剪文字图像. 没有与此数据集关联的词汇表.

SVT-P (SVT-Perspective)[22] 被用于文字识别,并且是一个不规则文字数据集. 主要由侧视文字组成,其图片来自于非正面拍摄的街景,因此很多图片都伴随强烈的视角扭曲. SVT-P 包含 639 张裁剪图片. 该测试集每张图片关联了一个 50 词的词汇表.

CUTE80[23] 专门用于评估弯曲文字识别的性能. 其包含 288 个裁剪的自然图像的测试集. 没有词汇表与此数据集相关联.

ICDAR 2015 (IC15)[24] 包含 2077 个裁剪图像,包括 200 多张不规则文字图片. 没有词汇表与此数据集相关联.

4.2 实现细节

(1) 网络结构:有关 FRN 和 AEN 的详细信息分别在表 1 和表 2 中给出. 解码器中 GRU 的隐藏单元数为 256. AEN 输出 37 个类别,包括 26 个字母,10 数字和 1 个代表 EOS 的符号.

(2) 模型训练:FRAEN 以端到端的方式进行训练. 训练数据由 Jaderberg 等人[25] 发布的 800 万张合成图像和 Gupta 等人[26] 发布的 600 万合成图像构成. 我们的实验中不使用任何像素级标签. 使用 Adadelta 自适应学习率调整的优化方法,我们在开始时将学习率设置为 1.0,每三个 epoch 之后降低 10 倍,批量大小设置为 256,训练完全耗费了 46 h 左右的时间.

(3) 实现:我们基于 PyTorch0.4 框架实现了我们的方法. 我们的实验中使用 NVIDIA RTX-2070 GPU,CUDA 10.0 和 CuDNN v7 后端,我们的模型使用 GPU 加速,所有图像尺寸都调整为 32×100.

4.3 FRAEN 在规则文字数据集上的性能

我们在常用规则文字数据集上进行评估,这些数据集中大多数测试样本是规则文字,其中有一小部分是不规则文字. 我们将本文方法与之前 9 种方法进行比较,结果如表 4 所示. FRAEN 在没有词汇表的模式下优于所有当前最好的方法.

4.4 在不规则文字上的识别结果

我们还在不规则文字数据集上进行了评估,在存在大量不规则文字的 SVT-P,CUTE80 和 IC15 三个测试集上进行测试. 结果如表 5 所示,FRAEN 表现优异.

对于 SVT-P 数据集,许多样本都是低分辨率和透视的. 具有 50 字词汇表的 FRAEN 的结果与 Liu 等人[27] 的方法的结果相同. 但是,FRAEN 在没有任何词汇表的情况下优于所有方法.

表 4　FRAEN 在规则文字测试集上的准确率

Tab. 4　Accuracy of the FRAEN on regular datasets.

方法	IIIT5K			SVT		IC03		IC13	
	50	1k	None	50	None	50	Full	None	None
Jaderberg 等人[16]	95.5	89.6	—	93.2	71.7	97.8	97.0	89.6	81.8
Shi，Bai 等人[18]	97.8	95.0	81.2	97.5	82.7	98.7	98.0	91.9	89.6
Shi 等人[8]	96.2	93.8	81.9	95.5	81.9	98.3	96.2	90.1	88.6
Le 和 Osindero[11]	96.8	94.4	78.4	96.3	80.7	97.9	97.0	88.7	90.0
Liu 等人[27]	97.7	94.5	83.3	95.5	83.6	96.9	95.3	89.9	89.1
Yang 等人[12]	97.8	96.1	—	95.2	—	97.7	—	—	—
Yin 等人[17]	98.7	96.1	78.2	95.1	72.5	97.6	96.5	81.1	81.4
Cheng 等人[9]	98.9	96.8	83.7	95.7	82.2	98.5	96.7	91.5	89.4
Cheng 等人[10]	99.6	98.1	87.0	96.0	82.8	98.5	97.1	91.5	—
本文方法	97.9	96.2	92.2	96.6	88.3	98.7	97.8	95.6	92.4

表 5　FRAEN 在不规则文字测试集上的准确率

Tab. 5　Accuracy of the FRAEN on irregular datasets.

方法	SVT-P		CUTE80		IC15
	50	Full	None	None	None
ABBYY[19]	40.5	26.1	—	—	—
Mishra 等人[20]	45.7	24.7	—	—	—
Wang 等人[6]	40.2	32.4	—	—	—
Shi 等人[8]	91.2	77.4	71.8	59.2	—
Yang 等人[12]	93.0	80.2	75.8	69.3	—
Liu 等人[28]	94.3	83.6	73.5	—	—
Cheng 等人[9]	92.6	81.6	71.5	63.9	66.2
Cheng 等人[10]	94.0	83.7	73.0	76.8	68.2
本文方法	94.3	86.7	76.1	79.5	72.7

4.5　FRAEN 的局限

为了公平比较和良好的可重复性，我们选择了广泛使用的训练数据集进行测试. 如图 6 所示，可以看出最后两张图像本文方法预测错误，因此，在场景文字背景复杂和文字弯曲角度太大时，本文方法可能会失效，因为其可能会错误地将复杂背景视为前景，从而影响预测结果. 上述实验均基于裁剪文字识别，没有文字检测器的 FRAEN 还不是端到端场景文字检测识别系统. 在更多应用场景中，不规则和多方向的文字对于检测和识别都具有很大的挑战性.

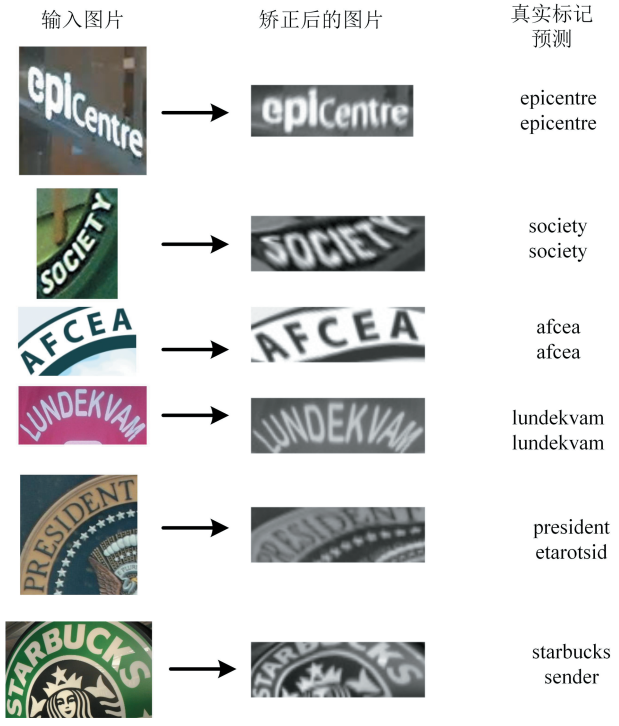

图 6　SVT-Perspective 和 CUTE80 数据集上的结果

Fig. 6　Results on SVT-Perspective and CUTE80

5　结　论

在本文中，我们提出了一个用于任意形状场景文字识别的带有灵活矫正功能的注意力增强网络. 本文方法分成两个阶段来解决不规则文字识别问题：文字矫正和文字识别. 首先，由矫正网络处理复杂的变形文字，将其矫正为更易识别的文字. 然后，使用基于相邻注意力权重的双向解码器的序列识

别网络来识别矫正后的图像并预测输出.我们在规则和不规则文字数据集上进行了大量实验,都表现出了优异的识别性能,尤其在不规则文字数据集上.将来,我们有必要扩展这种方法来处理任意方向和任意弧度的文字识别问题,由于文字和背景的多样性,这个问题更具挑战性.由于端到端文字识别性能的改进不仅取决于识别模型,还取决检测模型.所以,找到一种将FRAEN与场景文字检测器结合起来的正确有效方法也是值得研究的方向.

参考文献:

[1] 欧先锋,向灿群,郭龙源,等.基于Caffe深度学习框架的车牌数字字符识别算法研究[J].四川大学学报:自然科学版,2017,54:971.

[2] 张功国,吴建,易亿,等.基于集成卷积神经网络的交通标志识别[J].重庆邮电大学学报:自然科学版,2019,31:571.

[3] 魏超,范自柱,张泓,等.基于深度学习的农作物病害检测[J].江苏大学学报:自然科学版,2019,40:190.

[4] 王进,谢水宁,颉小凤,等.用于手写签名识别的演化超网络[J].重庆邮电大学学报:自然科学版,2018,30:399.

[5] 赵逸群,刘富,康冰.基于车牌检测的前方车辆识别方法[J].吉林大学学报:信息科学版,2019,37:168.

[6] Wang T, Wu D J, Coates A, *et al*. End-to-end text recognition with convolutional neural networks [C]// Proceedings of International Conference on Pattern Recognition. [s. l.]: IEEE, 2012: 3304.

[7] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. IEEE T Pattern Anal, 2017, 39: 2298.

[8] Shi B, Wang X, Lyu P, *et al*. Robust scene text recognition with automatic rectification [C]// Proceedings of Computer Vision and Pattern Recognition. [s. l.]: IEEE, 2016.

[9] Cheng Z, Bai F, Xu Y, *et al*. Focusing attention: towards accurate text recognition in natural images [C]// Proceedings of International Conference on Computer Vision. [s. l.]: IEEE, 2017.

[10] Cheng Z, Xu Y, Bai F, *et al*. AON: Towards arbitrarily-oriented text recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [s. l.]: IEEE, 2018: 5571.

[11] Lee C Y, Osindero S. Recursive recurrent nets with attention modeling for OCR in the wild [C]// Proceedings of Computer Vision and Pattern Recognition. [s. l.]: IEEE, 2016: 2231.

[12] Yang X, He D, Zhou Z, *et al*. Learning to read irregular text with attention mechanisms [C]// Proceedings of International Joint Conference on Artificial Intelligence. [s. l.]: Morgan Kaufmann, 2017.

[13] 李勤,师维,孙界平,等.基于卷积神经网络的网络流量识别技术研究[J].四川大学学报:自然科学版,2017,54:959.

[14] 龙彬,胡思才,郭峻铭,等.基于BP神经网络的网络小说排行预测[J].四川大学学报:自然科学版,2019,56:50.

[15] Zhu Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends [J]. Front Comput Sci, 2016, 10: 19.

[16] Jaderberg M, Simonyan K, Vedaldi A, *et al*. Deep structured output learning for unconstrained text recognition [C]// Proceedings of International Conference on Learning Representations. [s. l.]: [s. n.], 2015.

[17] Yin F, Wu Y C, Zhang X Y, *et al*. Scene text recognition with sliding convolutional character models [J]. arXiv, 2017, 1709: 1727.

[18] Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. IEEE T Pattern Anal, 2017, 39: 2298.

[19] Wang K, Babenko B, Belongie S. End-to-end scene text recognition [C]// Proceedings of International Conference on Computer Vision. [s. l.]: IEEE, 2011.

[20] Mishra A, Alahari K, Jawahar C. Scene text recognition using higher order language priors [C]// Proceedings of British Machine Vision Conference. [s. l.]: Springer, 2012.

[21] Karatzas D, Shafait F, Uchida S, *et al*. ICDAR 2013 robust reading competition [C]// Proceedings of the International Conference on Document Analysis and Recognition. New York: IEEE, 2013.

[22] Phan T Q, Shivakumara P, Tian S X, *et al*. Recognizing text with perspective distortion in natural scenes [C]// Proceedings of International Conference on Computer Vision. New York: IEEE, 2013.

[23] Risnumawan A, Shivakumara P, Chan C S, *et al*. A robust arbitrary text detection system for natural scene images [J]. Expert Syst Appl, 2014, 41: 8027.

[24] Karatzas D, Gomez-Bigorda L, Nicolaou A, *et al*. ICDAR 2015 competition on robust reading [C]// Proceedings of International Conference on Document Analysis and Recognition. [s. l.]: IEEE, 2015.

[25] Jaderberg M, Simonyan K, Vedaldi A, *et al*. Synthetic data and artificial neural networks for natural scene text recognition [C]// Proceedings of Advances in Neural Information Processing Deep Learn. [s. l.]: MIT Press, 2014.

[26] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images [C]// Proceedings of Computer Vision and Pattern Recognition. [s. l.]: IEEE, 2016.

[27] Liu W, Chen C, Wong K-Y K, *et al*. STAR-Net: A spatial attention residue network for scene text recognition [C]// Proceedings of British Machine Vision Conference. [s. l.]: Springer, 2016.

[28] Jaderberg M, Simonyan K, Vedaldi A, *et al*. Reading text in the wild with convolutional neural networks [J]. Int J Comput Vis, 2016, 116: 1.

引用本文格式：

中　文：徐富勇，余谅，盛钟松. 基于深度学习的任意形状场景文字识别[J]. 四川大学学报：自然科学版，2020，57：255.

英　文：Xu F Y, Yu L, Sheng Z S. Arbitrary shape scene text recognition based on deep learning [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 255.