文章编号：2096-1618(2025)01-0001-06

基于 CRNN 改进的中文街景文本识别技术

任 锐， 王晓娅， 文成玉
(成都信息工程大学通信工程学院，四川 成都 610225)

摘要：现实场景中存在图像扭曲、背景复杂、弯曲倾斜等不规则文字形状，提取其中的文字信息可提高图像的语义信息和帮助分析上下文，从而更好地理解场景图像。针对场景文本的复杂问题，提出基于 CRNN(卷积循环神经网络)改进的端到端场景文本识别技术。在卷积网络层提取特征，基于 GoogLeNet 改进的 inception 结构，加入多分支卷积层对多尺度特征的融合，其次融入注意力机制，在通道维度和空间维度加强特征联系，使局部特征拥有全局性。在循环网络层采用 Bi-LSTM(双向长短期记忆网络)加强字符之间的上下文联系进行序列预测，最后将预测序列传入 CTC(时序分类层)进行转录后序列输出。在 IIIT5K 数据集和百度中文街景数据集上的实验结果表明，该方法分别获得了95.3% 和91.1% 的准确率，证明其可靠性。

关 键 词：文本识别；卷积神经网络；注意力机制；双向长短期记忆
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2025.01.001

0 引言

随着互联网科技发展，人们获取信息的方式更加便捷，如图片、视频等。在自然场景图片中文字信息比其他图像信息拥有更多语义信息，可以获得如商店招牌、食品保质期、交通指示牌等关键信息。但由于场景图像的背景复杂、文字扭曲、图像倾斜等原因增加了提取文本信息的难度。

场景文本识别中，对于文档类的文字识别读取已经较为成熟，而复杂场景下的文本识别技术仍在发展中。Mishra 等[1]通过条件随机场结合图像底层和高级信息识别文字。该方法使用滑动窗口提取存在于复杂背景下的字符，并将这些从底层自下而上的信息与从字典模型中提取的高级信息相结合。如两个“o”从中间划开便容易分类成“x”，结合高级信息分析即可预测为两个“o”。Shi 等[2]提出的基于 CRNN 端到端的场景文本识别方法，在检测后的文本框里进行识别。采用 CNN+双向 LSTM+CTC 的方式，在卷积层将图片的一定宽度特征转化成宽度为 1 的特征向量，然后放入循环层对序列进行预测，采用双向的 LSTM 采集上下文信息，最后在 CTC 进行序列整合输出。该方法在 IIIT5K 数据集基于9 k个字符的字典下仅达到 78.2% 的准确率，且 LSTM 的顺序计算在运行时较为费时。Trinh 等[3]提出一种 CRNN(convolutional recurrent neural network)模型，在 ICDAR2013 数据集上实现89.6% 的准确率，实验后发现在中文场景数据集上表现很差。为解决传统 CRNN 模型在复杂场景下中文字符识别上的性能问题，本文做出了以下改进：

(1)采用改进的 CRNN 端到端的模型，将基于传统 CNN 的模型改进为基于 GoogLeNet 网络的 inception 结构的 CNN 层。

(2)在网络中对序列特征采用空间注意力和通道注意力(CMBA)和 inception 组合结构。

1 相关工作

1.1 CRNN 网络

传统的图像文本识别中，卷积神经网络使用非常广泛。在此基础上衍生出了许多改进网络，但基本组成结构只有 3 个模块：卷积层、循环层和转录层。传入的图像为 Image $\in R^{H\cdot W\cdot 3}$，场景图像通常为 RGB 三通道图片。在 CNN 层经过卷积后提取出多维特征，得到 $1\times1\times C$ 的 N 个特征序列，1×1 是尽可能地扩大感受野，C 为序列维度。随后，将 N 个序列放入 Bi-LSTM(双向长短期记忆网络)获取每个特征序列在整个序列里面的顺时和逆时的相关性。特征序列在经过 Bi-LSTM 处理得到序列的预测输出 $O=[O_1,O_2,\cdots O_N]$，$N\in R^L$，其中 L 是字典的字符数量。通过将 O 输入到转录层，进行转录输出，可以得到最终文本识别结果，并且这个结果不包含冗余信息。CRNN 的网络框架如图 1 所示。

收稿日期：2023-09-04
基金项目：四川省科技计划资助项目(2023YFS0422)

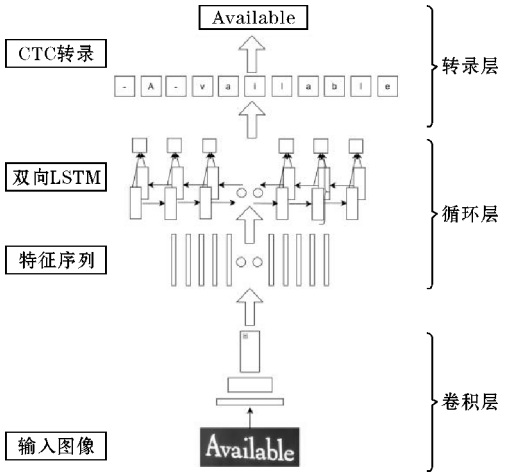

图 1 CRNN 网络框架

1.2 卷积神经网络层

传统卷积神经网络由卷积、池化和全连接等层组成。卷积操作通过一个窗口滤波器,在图像上滑动来提取局部特征。在同一维度采用相同的权重值矩阵进行运算,而后经过激活函数引入非线性特征得到最后的输出:

$$O=f(w^i \cdot x+b^i) \tag{1}$$

式中:O 是输出特征,w^i 是第 i 维的卷积层的参数,b^i 是卷积层的偏执计算量,$f(\cdot)$ 是激活函数。一般采用 ReLU、Sigmoid 或者 Tanh 增加非线性表达。

池化层主要用于减小特征图尺寸的同时保留重要的特征信息,在卷积层后进行降采样。

全连接层则是将局部信息与权重参数结合进行线性组合,再通过激活函数进行非线性变换。这种变换可以将低级特征转化为高级抽象的特征表示,有助于区分度特征的提取。

1.3 循环层

循环神经网络(recurrent neural network,RNN)是一种常用于处理序列数据的神经网络模型。在处理长序列信息时,反向传播容易出现梯度爆炸、梯度消失的问题,所以提出一些改进模型,如长短期记忆网络(long short-term memory,LTSM)。LTSM 通过引入特殊的门结构,选择性地控制信息的流动,从而允许网络选择性地保留或忘记信息,以解决梯度问题,增强序列信息的长期依赖的建模能力。

在 LSTM 中,包含 3 种门:输入门,如式(2);遗忘门,如式(3);输出门,如式(4)。

$$I_t=\sigma(X_tW_{xi}+H_{t-1}W_{hi}+b_i) \tag{2}$$

$$F_t=\sigma(X_tW_{xf}+H_{t-1}W_{hf}+b_f) \tag{3}$$

$$O_t=\sigma(X_tW_{xo}+H_{t-1}W_{ho}+b_o) \tag{4}$$

式中,σ 是 Sigmoid 激活函数,X_t 为本单元的输入,H_{t-1} 为上个单元的隐藏状态,W_x 为输入的权重,W_h 为隐藏层的权重,b 为偏执。通过本单元输入以及上个单元的隐藏层输出计算出 3 个门的具体值:

$$\tilde{C}_t=\tanh(X_tW_{xc}+H_{t-1}W_{hc}+b_c) \tag{5}$$

$$C_t=F_t \cdot C_{t-1}+I_t \cdot \tilde{C}_t \tag{6}$$

$$H_t=O_t \cdot \tanh(C_t) \tag{7}$$

式中,“·”为点乘操作,通过 Tanh 激活函数对上一个单元的隐藏层和当前单元的输入进行处理,得到单元 c 的候选值。然后通过更新门与单元 c 的候选值的计算,遗忘门和上个单元状态值的计算决定当前的状态值,最后输出门与当前状态值的计算得出当前隐藏层的输出值。从这个结构可以看出,更新门控制是否将候选值存储为记忆,而遗忘门控制是否遗忘上一个单元状态值或继续保留。

1.4 转录层

传统 CRNN 模型的转录层使用的是 CTC(connectionist temporal classification)接时序分类,是一种用于处理序列数据的时序分类方法。通过在标签序列中引入空白符号,使用反向传播算法进行训练,无须对齐标签和输入序列简化了训练过程。

CTC 损失函数是基于对所有可能路径的概率进行求和,然后通过反向传播算法进行优化以更新模型参数,公式为

$$p(L|x)=\sum_{\pi\in\beta^{(-1)}(L)}p(\pi|x) \tag{8}$$

式(8)表示在给定输入序列 x 的情况下,通过计算所有可能的输出序列 π 的概率 $p(\pi|x)$,并对符合条件的输出序列进行求和操作,得到输出序列 L 的概率 $p(L|x)$。其中,$\beta^{(-1)}(L)$ 表示与目标输出序列 L 对应的所有可能输出序列的集合。

通过最小化损失函数,CTC 可以学习到将输入序列映射到正确输出序列的模型参数,从而实现对时序分类任务的准确预测。

2 场景图像文本识别算法

2.1 GoogLeNet 的 inception 结构改进

CRNN 网络面对中文场景的图像文本识别时,效果通常不好。相对于英文字符,中文字符具有更高的复杂性和多样性,存在大量的复杂字形和变体,使特征提取更加困难。其次在垂直方向上,中文字符的密集度较英文字符更高,而传统 CRNN 网络对于垂直方向特征提取有一定局限性。

基于上述问题,将传统卷积网络的核心结构用改进的 GoogLeNet 的 inception 结构代替,利用多分支融合卷积以及非对称卷积提升特征提取性能。在原 GoogLeNet 网络的基础上选择去除辅助分类器,并将 inception 结构进行适合应用场景的改动。

首先将改进 GoogLeNet 的架构分为 inception1、inception2 和 inception3。

Inception1 为原文原始结构,由于图像大小原因只是将其中 7×7 的卷积替换成 5×5 的卷积,如图 2 所示。

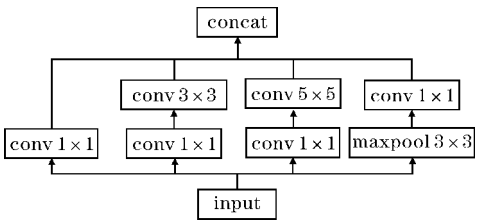

图 2　inception1 结构

图 2 结构在同一层级融合了多个尺度信息,通过多个不同大小的卷积核并行处理输入信息,从而捕捉不同尺度的特征,这对处理不同大小和形状的图像特征有较大帮助,增强了网络的特征表达能力。并通过不同大小的卷积核带来不同的感受野,融合这些输出可以扩大整个网络的感受野,感知更多的上下文信息。其次 inception 结构带来了 1×1 的卷积,这个卷积能在变化维度时减少计算量和最大化保留特征。

改进后的 inception2 结构如图 3 所示。

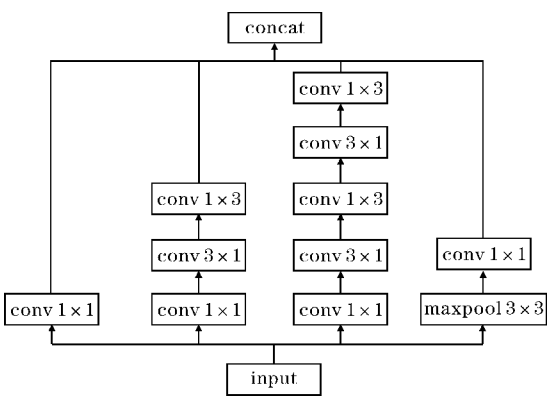

图 3　inception2 结构

inception2 结构在 inception1 的基础上将 5×5 的卷积核变成两层 3×3 的卷积叠加,再将 3×3 的卷积分成两组 3×1 和 1×3 的非对称卷积。此举增加了网络的深度,对于特征的深层信息捕捉能力更强。其次,针对中文文字的特殊性引入非对称卷积,加强了对纵向(如偏旁)的特征提取以及横向文字结构的特征提取,提高了横向和纵向的特征捕捉能力。在不改变感受野的条件下降低了网络的计算量。

改进后的 inception3 结构如图 4 所示。

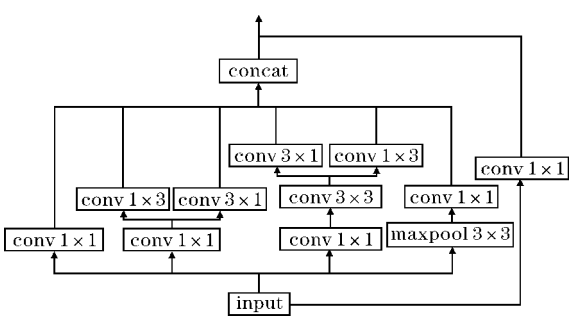

图 4　inception3 结构

inception3 结构将 inception2 的串行非对称卷积设计成并行。这一步不仅扩大了网络的宽度,更着重在提取深层被压缩的特征时,减少横向及纵向特征的信息瓶颈,对于同一输入特征图能实现部分参数共享。其次,引入残差结构[7],用 1×1 的卷积核作为捷径分支的操作,解决在后层的网络出现梯度消失或梯度爆炸现象,起到特征复用的作用。

通过宽网络的设计,可以获得特征的多尺度感受野,提升对上下文信息的理解。其次,通过同一层级不同卷积的相同输入特征图,共享部分参数,减少模型的参数数量,降低了过拟合风险,提升学习效率。通过将多个不同尺寸卷积核的输出连接在一起,形成特征重组,使网络更好理解图像特征内容。在网络信息传递时,因为卷积池化会压缩信息,多尺度融合也能使信息瓶颈得到改善,使得网络更好地保留和传递特征。

2.2　注意力机制

注意力[8]是在深度学习中用于对输入进行选择性处理和增强特征感知力的技术。在 CRNN 网络中通过引入空间注意力模块和通道注意力模块,从而增强网络的表达能力,将权重更多放置于文字特征。

空间注意力结构如图 5 所示。

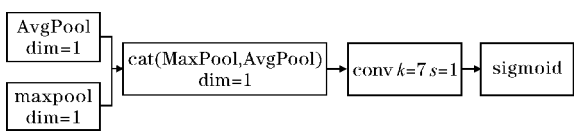

图 5　空间注意力结构

$$S(x)=\mathrm{sigmoid}(k^{7\times7}(\mathrm{AvgPool}(x)\oplus\mathrm{MaxPool}(x))) \tag{9}$$

式(9)表示空间注意力机制使用最大池化和平均池化对输入特征的通道维度进行压缩,得到通道数为 1 的输出特征。然后在维度下进行拼接,再经过一个 7×7的卷积核的卷积后,放入 Sigmoid 激活函数处理。该模块将通道维度压缩为 1 而空间维度不变,以关注特征的位置信息。

通道注意力结构如图 6 所示。

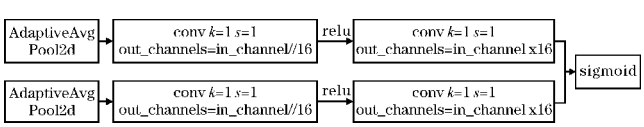

图 6　通道注意力结构

$$C(x)=\mathrm{sigmoid}(\mathrm{MLP}(\mathrm{AvgPool}(x))+\mathrm{MLP}(\mathrm{MaxPool}(x))) \tag{10}$$

式(10)表示通道注意力机制首先使用自适应平均池化和自适应最大池化将特征空间尺寸设为 1×1,然后定义一个两层卷积层加 ReLU 激活函数的多层感知机模块(MLP)进行非线性变换,将两个不同的输出特征进行空间维度相加,经过 sigmoid 函数激活得到通

道注意力 $C(x)$,以关注特征的重要维度信息。

这两个模块可插在网络的任意位置,通过与输入特征 x 相乘将权值加入进去,输入到下一层,即

$$x=x\cdot S(x) \tag{11}$$

$$x=x\cdot C(x) \tag{12}$$

一般来说,将空间注意力放在网络的浅层或中间层可更好地提高局部细节的捕捉能力,而通道注意力适用于关注全局语义信息,将其放置在网络的深层可以获得较为全面的语义相关性表示。

2.3 文本特征提取网络

在传统卷积神经网络里,通常是追求更深的网络来提供更大的模型参数容量以及表达能力,有助于对更复杂特征的捕捉。应用于图像分类时,因为更深层以及更多维的特征可以学习到更高级的抽象表示从而提高模型的准确性。当处理图像文本识别时,由于文本相对于分类来说拥有更明显的几何特征以及字符间有空间间隔,可以通过对局部区域的分析而不需要太过于考虑上下文的关系,从而减少对网络深度的需求。

经验证后得出对于中文文本特征的提取,需要加大网络的宽度并应用非对称卷积以获得多尺度特征。所以,本文选取 GoogLeNet 的 inception 为基础框架,在其基础上进行改进以适用于中文街景图像文本识别的应用场景,形成新的卷积神经网络框架。通过多次实验,图 7 框架结构在当前情况下取得最优结果。

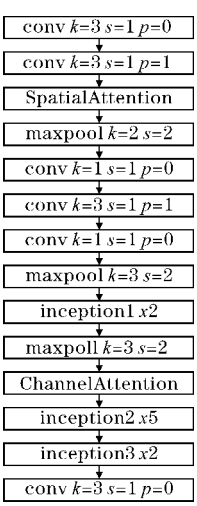

图 7 网络框架

这里的卷积操作默认包含了卷积、批归一化和 ReLU 激活函数。k 为卷积核的大小,s 是操作的步进,p 是边缘填充大小。

3 实验结果与分析

3.1 数据集

实验采用 IIIT5K 的英文场景图片数据集和百度的中文街景数据集(飞浆常规赛)。IIIT5K 包含 5000 张图片,实验将其分成 4000 张训练集和 1000 张测试集。百度的中文街景数据集包含 50000 张图片,实验将其分成 40000 张训练集和 10000 张测试集。实验时,将对英文场景和中文场景进行性能评估。

3.2 评价指标

在场景文本识别中,一般用字符串编辑距离来评估识别结果与标签的相似程度,衡量两个字符串相等过程中所需的插入、替换和删除操作次数。每变动一次,距离就会加 1。公式为

$$\mathrm{ACC}=\frac{N-\mathrm{Ins}-\mathrm{Sub}-\mathrm{Del}}{N} \tag{13}$$

式中,N 为文本标签所对应的数量,Ins 为插入的字符数量,Sub 为替换的字符数量,Del 为删除的字符数量。

3.3 实验环境及数据处理

实验在 Windows 操作系统下进行,python 版本为 python3.8,GPU 型号为 NVIDIA RTX4070 8GB,CUDA 版本为11.7。算法在 Pytorch2.0.1 框架下实现,采用 Adadelta 算法,自适应设置学习率。

因为数据集的图片大小尺寸不一致,网络在处理过程会有偏差,所以将图片统一成相同的尺寸。将高度固定后成比例放大或缩小图片,然后在宽度上进行裁剪或者补空白操作,最终将图片的大小设置为 960×48。

3.4 实验结果分析

在百度的中文街景数据集上对于原始的 CRNN 模型进行验证,修改部分卷积层和池化层参数后,得到如图 8 所示的结果。

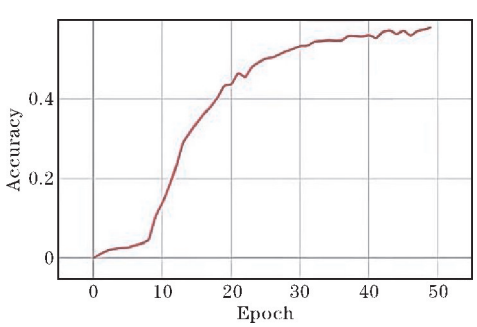

图 8 CRNN 模型验证结果

在本文创新中,首先将原始 CRNN 模型的主干替换为改进的 inception 结构。经过 30 轮学习后,结果如图 9 所示,改进后的网络准确率达83.6%。可以看出,改进后的 CRNN 模型网络的收敛更快,准确率有较大提升。由此可见,图像文本识别上,不是网络越深越好。对于场景图片文字和文本的视觉表达通常需要包含大量细节和上下文信息,网络需要学习更多的自由度用于学习更复杂更丰富的特征。而采用多卷积的宽

网络可以提供多尺度的信息处理能力,对于文字在场景的变化也能快速捕捉,增加网络的鲁棒性。所以,在将改进的 inception 结构引入后,网络对于文字特征的捕捉能力加强,在前期训练过程中会加速网络的收敛,反向传播时更有利于梯度下降。

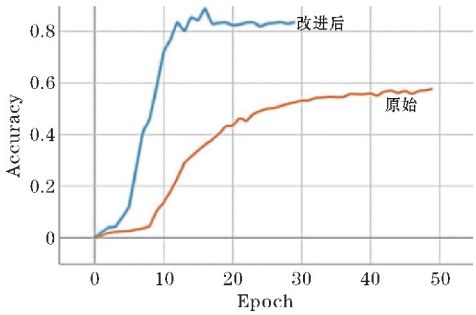

图 9　改进的 CRNN 模型结果

然后,在其上加入注意力作对比。将空间注意力放置于网络浅层,通道注意力放置网络中后层。最后,得到图 10 的对比图。加入注意力后的网络准确率有一定的提升。可知注意力机制在网络训练时能够帮助模型在图像中定位并关注与文本相关的区域与相关的维度特征,提高模型对文本的理解识别能力,进而提高准确性和鲁棒性。最终达到91.1% 的准确率。

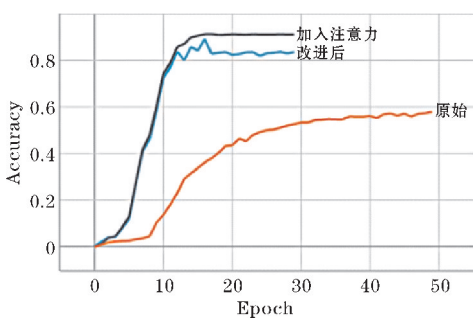

图 10　改进的文本识别网络结果

在 IIIT5K 的英文场景文本小数据集上验证改进的文本识别网络,识别准确率如图 11 所示,准确率达 95.3%。可见,在文本特征相对简单的英文场景下,可取得更好的处理结果。

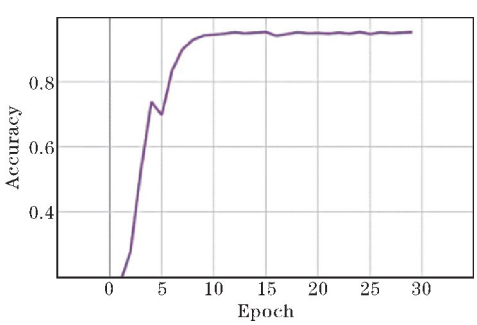

图 11　改进网络在 IIIT5K 上的准确率

在 IIIT5K 数据集上,本文与其他模型的准确率的对比如表 1 所示。

表 1　实验结果对比

模型	数据集(无词典)	识别准确率/%
CRNN[1]	IIIT5K	78.2
闫郁瑾[4]	IIIT5K	87.20
丁宇[5]	IIIT5K	92.10
熊炜[17]	IIIT5K	93.90
XinTang[18]	IIIT5K	95.10
本文模型	IIIT5K	95.30

可以看出,本文提出的方法在没有对图片做出预处理和无词典情况下准确率较其他模型有一定的提高。由此可知,本模型在原模型基础上的改进,对于场景文本特征提取有较好的效果。

4　结束语

针对场景图像的文本识别提出了基于 GoogLeNet 的 inception 结构改进方法,在特征提取时并未一味追求网络的深度而尽量拓宽网络的宽度。结果表明,对多尺度特征融合的结构以及添加注意力可以增强网络对背景复杂的文本内容特征提取能力。在接下来的工作中,将继续尝试循环层和转录层的改进工作,或者引入词典以及语义分析等后处理,增强文本识别性能。

参考文献:

[1]　Mishra A, Alahari K, Jawahar C. Top-down and bottom-up cues for scene text recognition[C]. Top-down and bottom-up cues for scene text recognition. 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2012:2687-2694.

[2]　Shi B, Xiang B, Cong Y. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition[J]. Proceedings of IEEE Transactions on Pattern Analysis & Machine Intelligence,2016,39(11):2298-2304.

[3]　Trinh Tan Dat, Le Tran Anh Dang, Nguyen Nhat Truong. An improved CRNN for Vietnamese Identity Card Information Recognition[J]. COMPUTER SYSTEMS SCIENCE AND ENGINEERING. 2022,40(2):539-555.

[4]　闫郁瑾. 基于 CRNN 的自然场景文字识别算法研究[D]. 西安:西安电子科技大学,2020.

[5] 丁宇. 基于深度学习的自然场景文字识别研究[D]. 青岛:山东科技大学,2020.

[6] Christian Szegedy, Wei Liu, Yangqing Jia. Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:7-12.

[7] He Kaiming, Zhang Xiangyu, Ren. Shaoqing Deep Residual Learning for Image Recognition[C]. CoRR. Volume. abs,2015:45-49.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems,2017:30-30.

[9] 薛晨兴,张军,邢家源. 基于 GoogLeNet Inception V3 的迁移学习研究[J]. 无线电工程,2020,50(2):118-122.

[10] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]. Proceedings of 31st AAAI Conference on Ar-tificial Intelligence, AAAI, 2017:4278-4284.

[11] Hassan Ehtesham, Lekshmi V L. Attention Guided Feature Encoding for Scene Text Recognition[J]. Journal of Imaging,2022,8(10):276-276.

[12] Kantipudi MVV Prasad, Kumar Sandeep, Kumar Jha Ashish. Scene Text Recognition Based on Bi-directional LSTM and Deep Neural Network[J]. Computational Intelligence and Neuroscience, 2021,11(5):13-15.

[13] 陈鹏,李鸣,张宇,等. 一种端到端的自然场景文本检测与识别模型[J]. 测控技术,2022,41(7):17-22.

[14] Yousef Mohamed, Bishop Tom 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,2020:14710-14719.

[15] Zuo Lingqun, Sun Hongmei, Mao Qichao. Natural Scene Text Recognition Based on Encoder-Decoder Framework[J]. IEEEAccess,2019,7:62616-62623.

[16] 吴启明,宋雨桐. 基于 YOLOv3 与 CRNN 的自然场景文本识别[J]. 计算机工程与设计,2022,43(8):2352-2360.

[17] 熊炜,孙鹏,强观臣. 基于注意力机制的自然场景图像中文本识别方法及系统[P]. 中国专利,202310120821.8,2023-2-13.

[18] Xin Tang, Yongquan Lai, Ying Liu. Visual-Semantic Transformer for Scene Text Recognition [J]. arXiv preprint arXiv,2021,12(8):56-61.

Improved Chinese Street View Text Recognition Technology based on CRNN

REN Rui, WANG Xiaoya, WEN Chengyu

(College of Communicating Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: In real-world scenarios, there are complexities such as image distortion, background clutter, bending, and tilting that can cause irregular text shapes. Extracting textual information from these images can enhance their semantic content and help analyze the context, thus better-facilitating understanding of the scene. To address these challenges in scene text recognition, an end-to-end text recognition technique based on CRNN (Convolutional Recurrent Neural Network) is proposed. In the convolutional network layer, an improved inception structure based on GoogLeNet is used to extract features. This structure incorporates multi-branch convolutional layers for the fusion of multi-scale features. Additionally, an attention mechanism is incorporated to enhance feature correlation in both the channel and spatial dimensions, giving local features a global perspective. In the recurrent network layer, Bi-LSTM (Bidirectional Long Short-Term Memory) is employed to strengthen the contextual relationships between characters for sequential prediction. Finally, the predicted sequence is fed into CTC (Connectionist Temporal Classification) for post-transcription sequence output. Experimental results on the IIIT5K dataset and Baidu's Chinese Street View dataset demonstrate the reliability of this approach, with accuracy rates of 95.3% and 91.1% respectively.

Keywords: text recognition; convolutional neural network; attention mechanism; bi-directional long and short-term memory