文章编号：1003－1251(2024)04－0027－05

# 面向文本识别的 CRNN 模型的改进

吕艳辉，刘明鑫

（沈阳理工大学 信息科学与工程学院，沈阳 110159）

**摘　要：**复杂场景下文本识别因阴影、残缺、模糊、虚化等因素会出现识别精度下降问题。鉴于此，提出一种基于特征融合与双向简化门结构的 CRNN 模型。首先引入特征融合机制改进卷积神经网络(CNN)模型，利用特征金字塔结构，多加一条自底向上的路径，将低层特征与高层特征融合在一起，以保留更多低层细节特征，提高场景文本识别精度；其次通过合并遗忘门与输入门，得到结构更简单、计算量和参数量更少的简化门结构替换长短期记忆(LSTM)网络改进循环神经网络(RNN)模型部分；最后设计消融实验验证改进后模型的有效性。三个数据集的测试结果表明：在 ResNet50 做主干网络时，与原始模型相比，改进后模型准确率提升了 1.5% 以上；在 MobileNetV3 做主干网络时，准确率提升了 1.4% 以上。

**关　键　词：**特征融合；长短期记忆网络；简化门结构

**中图分类号：**TP391.41　　**文献标志码：**A　**DOI：**10.3969/j.issn.1003－1251.2024.04.005

## Improvement of CRNN Model for Text Recognition

LÜ Yanhui，LIU Mingxin

（Shenyang Ligong University，Shenyang 110159，China）

**Abstract：** In complex scenarios，text recognition may experience a decrease in recognition accuracy due to factors such as shadows，imperfections and blurring. In view of this，a CRNN model based on feature fusion and bidirectional simplified gate structure is proposed. Firstly，a feature fusion mechanism is introduced to improve the CNN model. Utilizing the feature pyramid structure，an additional bottom-up path is added to fuse low-level features with high-level features，in order to retain more low-level detailed feature information and improve the accuracy of scene text recognition. Secondly，by merging forgetting gates and input gates，a simplified gate structure with less computation and parameter complexity is used to replace LSTM to improve the RNN model. Finally，ablation experiments are conducted to verify the effectiveness of the improved CRNN model. By testing three datasets the experimental results show that when ResNet50 is used as the backbone network，the accuracy of the proposed model is improved by more than 1.5% compared to the original model；when using MobileNetV3 as the backbone network，the accuracy is improved by over 1.4%.

**Key words：** feature fusion；long short-term memory network；simplified gate structure

应用场景中的文本图像蕴含丰富的文本信息，识别场景图像中的文本具有重要的价值。图像中的文本往往存在对比度低、文字倾斜模糊、字体种类繁多或图片质量不佳等问题，使复杂场景

**收稿日期：**2023－10－12

**基金项目：**辽宁省教育厅高等学校基本科研项目(JYTMS20230192)

**作者简介：**吕艳辉(1971—)，女，教授，博士，研究方向为计算机视觉、人工智能。

下的文本识别较为困难。随着深度学习技术的发展,其网络结构更加简单灵活且便于训练,逐渐被应用到场景文本的识别中。

基于深度学习的文本识别主流算法主要有两种,分别是基于 CTC[1] 的算法和基于 Sequence2Sequence[2] 的算法,两者区别在于解码阶段。基于 Sequence2Sequence 的算法由编码器和解码器两部分组成,其中编码器负责对输入内容的理解,并转化为隐秘状态,解码器负责对理解后的向量进行处理并获得输出。卷积循环神经网络(CRNN)[3] 是基于 CTC 的典型算法,特征提取部分使用主流的卷积结构,常用的有 ResNet[4]、MobileNet[5]、VGG[6] 等。

CRNN 分为卷积神经网络(CNN)[7] 与循环网络(RNN)[8] 两部分,其中 CNN 作为底层的骨干网络,用于从输入图像中提取特征序列[9]。文献[10]在 CRNN 基础上引入双向长短期记忆网络(LSTM)[11] 增强上下文建模,以有效提取图片的上下文信息,最后将输出的特征序列输入到 CTC 模块中。该方法已被验证有效并广泛应用在文本识别任务中。然而,因 LSTM 复杂的多门结构产生了庞大的计算量,增加了训练难度。

综上,本文以 CRNN 为基础研究面向文本识别的方法。首先引入特征融合机制改进 CRNN 模型的 CNN 部分,借鉴特征金字塔结构,多加一条自底向上的路径,使底层信息更容易传递到高层顶部,底层特征与高层特征融合可以保留更多低层细节特征,提高场景文本识别精度。其次改进 LSTM,通过合并遗忘门与输入门得到结构更简单、计算量和参数量更少的简化门结构,以此改进 CRNN 模型的 RNN 部分,提高文本识别性能。

# 1 基于特征融合与双向简化门结构的 CRNN 模型

## 1.1 特征融合的引入

在 CRNN 模型中,为提高识别精度与获得更加抽象的特征,特征提取部分会倾向于选择包含更多卷积层的主干网络,以获取到更高维度的细节特征。但是随着网络层数增加,容易丢失图像细节特征。因此,融合不同尺度的特征是提高图像识别效果的重要手段。在提取图像特征时,低层特征分辨率较高,但噪声较多;高层特征虽然具有较强的语义信息,但是分辨率低,对细节的感知能力较差。因而改进模型提升性能的关键是将高低层特征高效融合。

本文借鉴特征金字塔结构[12] 在 CRNN 模型的 CNN 部分引入特征融合机制,该机制依靠路径聚合网络(path aggregation network,PANet)[13] 实现,通过自底向上的路径增强,利用准确的低层定位信号增强整个特征层次,从而缩短低层与高层特征之间的信息路径。改进的模型结构如图 1 所示,图 1 中 b 部分是引入的特征融合路径,选用 ResNet50 作为主干网络。

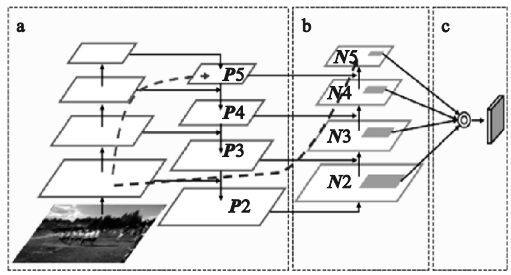


**图 1 引入特征融合后的模型结构**

**Fig. 1 The model structure by introducing feature fusion**

图 1 中 a 部分采用自顶向下的路径方法,得到 $P5$、$P4$、$P3$、$P2$,共计 4 组特征图。当图像经过神经网络的各层时,特征的复杂度不断增加,同时图像的空间分辨率相应降低。顶部信息向下逐层传递的方法计算量比较大,本文在此基础上增加一条自底向上的路径,如图 1 中 b 部分所示,得到 $N2$、$N3$、$N4$、$N5$ 等 4 组新的特征图,使底层信息更容易传递到高层,同时使用从底层到顶层的横向连接以缩短路径。

图 1 中的两条虚线部分为自顶向下模块和自底向上模块中添加的跨越多层的路径,可以缩短浅层特征传递到顶层所经过的路径,解决了经过多层传递导致浅层特征信息丢失严重的问题。

每层采样后都对得到的特征信息进行检测,最终融合所有层的检测信息以得到一个更加完整且高度接近于原图的数据,如图 1 中 c 部分所示。

## 1.2 双向简化门结构的提出

本文对 LSTM 进行改进,提出简化门结构。其结构是一个双门系统,把遗忘门与输入门合并成一个新的门,称为迭代门。不同于 LSTM 的第三个输出门,简化门结构另外一个门称之为重置门。相较于 LSTM 三门结构的众多参数,简化门结构更加简单,能够更好地解决长期依赖带来的梯度消失问题,其结构如图 2 所示。

### 1.2.1 重置门

图 2 中 $r_t$ 为重置门,负责统计当前时刻之前

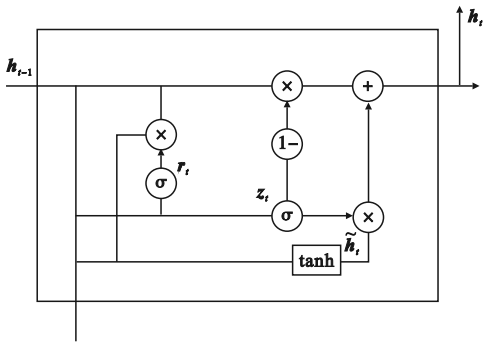


**图 2　简化门结构图**

**Fig. 2　The structure diagram of simplified gate**

所需要遗忘的信息数量,计算公式为

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{1}$$

式中:$\delta$ 为 sigmoid 函数,将数据变为[0,1]范围内的数值;$\boldsymbol{h}_{t-1}$为上一时刻的隐秘状态,包含节点之前的数据信息;$\boldsymbol{x}_t$ 是当前时刻的输入信号;$\boldsymbol{W}_r$ 为重置门的权重矩阵。

进一步可得当前时刻候选隐秘状态 $\tilde{\boldsymbol{h}}_t$ 为

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{\tilde{h}_t} \cdot [\boldsymbol{r}_t \cdot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{2}$$

式中:$\boldsymbol{W}_{\tilde{h}_t}$是当前时刻候选隐秘状态的权重矩阵;tanh 函数将数据变为[−1,1]范围的数值。

### 1.2.2　迭代门

图 2 中 $\boldsymbol{z}_t$ 为迭代门,$t$ 时刻迭代门控制 $t-1$ 时刻的隐秘状态 $\boldsymbol{h}_{t-1}$和 $t$ 时刻的隐秘状态 $\boldsymbol{h}_t$,包括监测有多少信息进入到当前时刻 $t$ 的隐秘状态 $\boldsymbol{h}_t$,迭代门数值越小,采集的信息越少,反之信息越多,越有利于监控时间序列中长期的依赖关系,其计算公式如式(3)所示,其中 $\boldsymbol{W}_z$ 是迭代门的权重矩阵。

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3}$$

当前时刻 $t$ 的隐秘状态 $\boldsymbol{h}_t$ 计算公式为

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \cdot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \cdot \tilde{\boldsymbol{h}}_t \tag{4}$$

在 LSTM 与简化门结构中分别输入以下相同参数,如表 1 所示。

**表 1　网络的输入参数**

**Table 1　Input parameters of the network**

| 参数 | 数值 |
|---|---|
| 时间步 | 28 |
| 每个时间步的特征长度 | 28 |
| 隐藏神经元个数 | 100 |
| 输出长度 | 10 |

表 2 给出了针对不同数据集的操作时间。

由测试结果可知,相比 LSTM,简化门结构少了一个门控结构,参数量减少,运行速度更快。

**表 2　针对不同数据集的操作时间**

**Table 2　Operation time for different datasets**　h

| 数据集 | 操作 | 简化门结构 | LSTM |
|---|---|---|---|
| Nottingham[14] | 训练 | 2.79 | 3.08 |
| | 测试 | 3.20 | 3.23 |
| MuseData[15] | 训练 | 5.06 | 5.18 |
| | 测试 | 5.99 | 6.23 |
| Piano-midi[16] | 训练 | 4.93 | 6.49 |
| | 测试 | 8.82 | 9.03 |

单向简化门结构在处理特征向量时只能利用当前时刻之前的上下文信息,割裂了特征向量上下文的整体联系。为使特征向量含有充足的时序信息,本文采用双向的简化门结构,如图 3 所示。将提取到的特征序列输入到双向简化门结构中,使其带有时序信息,最终将识别到的字符拼接成完整且正确的语义信息。

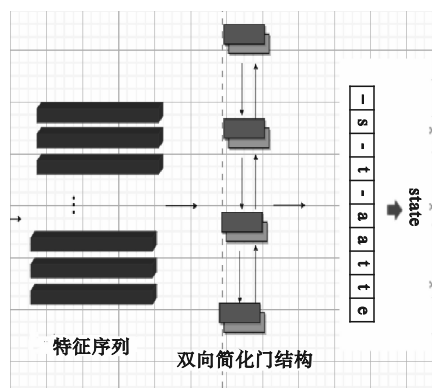


**图 3　双向简化门结构改进模型图**

**Fig. 3　The improved model diagram based on bidirectional simplified gate structure**

## 1.3　改进后的 CRNN 模型

通过整合对 CRNN 模型的两种改进方法,设计出基于特征融合和双向简化门结构的模型,其结构如图 4 所示。

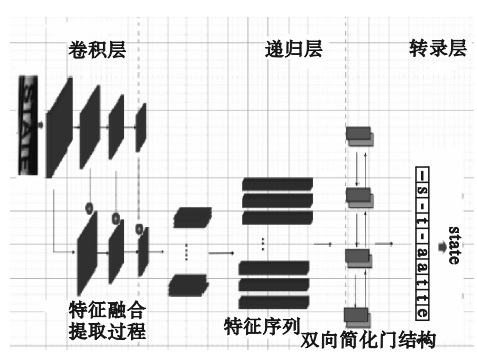


**图 4　基于特征融合与双向简化门结构改进的 CRNN 模型**

**Fig. 4　The improved CRNN model based on feature fusion and bidirectional simplified gatestructure**

首先在 CNN 部分加入特征融合模块,模型不仅提取到最后一层特征,在卷积层结合每一层的特征;在 RNN 部分使用简化门结构,结构简单、参数量与计算量较小。

### 1.4 数据集

本文模型首先在 SynthText[17] 和 Synth90K[18] 数据集上进行预先训练,SynthText 主要应用在自然场景领域中的文本检测识别研究,Synth90K 数据集主要提供合成场景文本图像;然后在以下 3 个公共数据集上进行测试。

ICDAR2003(IC03)包含 867 张图,用于阅读比赛;ICDAR2013(IC13)包含 251 个水平文字的完整场景图像和 860 张有单词被裁剪的图像;ICDAR2015(IC15)数据集的图像数据由 Google 提供,作为 ICDAR 举办的场景文本检测比赛的官方数据集,其中包括了 1 000 多张用于训练的图像和 500 多张用于测试的图像,大部分图像来自 IC03。

### 1.5 算法评估标准

文本识别模型的评估标准主要采用文本识别结果的准确率。当文本的人工标注结果与文本识别模型的识别结果相同时,可以认为文本识别结果为一次正确的识别,否则为错误的识别结果。使用所有文本识别算法识别出正确的文本数除以人工标注数据的总数得到最终的文本识别准确率 $p$,计算公式为

$$p = \frac{TP}{TP + FP} \tag{5}$$

式中:$TP$ 代表识别正确的文本数;$FP$ 代表识别错误的文本数。

## 2 实验结果与分析

### 2.1 实验结果

为提高识别精度与获得更加抽象的特征,特征提取部分选择包含更多卷积层的 ResNet50 和 MobileNetV3 作为主干网络。在数据集 IC03、IC13、IC15 上通过消融实验验证本文所提模型的可行性,实验结果如表 3、表 4 所示,表中√表示使用对应的改进,×表示未使用对应的改进,均未使用改进的是原始模型 CRNN,均使用改进的是本文所提模型。

由表 3、表 4 可以看出,在主干网络分别使用 ResNet50 与 MobileNetV3 时,与原始的 CRNN 模型相比,无论是引入特征融合还是双向简化门结构,准确率都有所提升。使用 ResNet50 主干网络时,与原始的 CRNN 模型相比,本文所提模型在三个数据集上的准确率均提升了 1.5% 以上;使用 MobileNetV3 主干网络时,准确率均提升了 1.4% 以上。

**表 3 不同模型在数据集 IC03、IC13、IC15 上的准确率(ResNet50)**

**Table 3 Accuracy of different models on datasets IC03, IC13, and IC15 (ResNet50)** %

| 特征融合 | 双向简化门结构 | IC03 准确率 | IC13 准确率 | IC15 准确率 |
|---|---|---|---|---|
| × | × | 92.7 | 88.6 | 77.3 |
| √ | × | 93.7 | 89.5 | 78.1 |
| × | √ | 93.1 | 89.0 | 77.9 |
| √ | √ | 94.3 | 90.3 | 78.8 |

**表 4 不同模型在数据集 IC03、IC13、IC15 上的准确率(MobileNetV3)**

**Table 4 Accuracy of different models on datasets IC03, IC13, and IC15 (MobileNetV3)** %

| 特征融合 | 双向简化门结构 | IC03 准确率 | IC13 准确率 | IC15 准确率 |
|---|---|---|---|---|
| × | × | 89.5 | 85.3 | 74.6 |
| √ | × | 90.2 | 86.2 | 75.4 |
| × | √ | 90.0 | 85.6 | 75.1 |
| √ | √ | 90.9 | 87.1 | 76.2 |

### 2.2 文本识别效果展示

图 5、图 6 给出了使用本文改进后的 CRNN 模型对场景文本图像识别的效果,可以看出,使用本文改进后的 CRNN 模型能够正确识别图像中的文本。


**图 5 文本识别效果**

**Fig. 5 The text recognition result**


**图 6　文本识别效果**

**Fig. 6　The text recognition result**

## 3　结论

本文以 CRNN 为基础,研究面向文本识别的方法。首先引入特征融合机制,将低层特征与高层特征融合在一起,提升了场景文本识别精度。其次,提出双向简化门结构,使输出的特征向量能够包含更加丰富的上下文信息。最后,将本文所提模型分别使用主干网络 ResNet50 和 MobileNetV3 在三个数据集上进行测试,实验结果表明,与原始模型 CRNN 相比,本文所提模型的准确率均有一定提升。

**参考文献**(References):

[1] 齐秀芳,吴陈. 不规则场景文本的识别方法[J]. 软件导刊,2022,21(6):200 – 204.
QI X F, WU C. Recognition method of irregular scene text [J]. Software Guide, 2022, 21(6): 200 – 204. (in Chinese)

[2] 韩珊珊,王升辉,万丽莉. 一种面向新闻文本的生成式中文摘要生成模型[J]. 中国传媒大学学报(自然科学版),2023,30(3):24 – 30.
HAN S S, WANG S H, WAN L L. A novel generative Chinese summarization model geared towards news text generation[J]. Journal of Communication University of China(Science and Technology), 2023, 30(3): 24 – 30. (in Chinese)

[3] 张少宇. 基于人工智能机器学习的文字识别方法研究[J]. 电脑编程技巧与维护,2022(9):154 – 156,176.
ZHANG S Y. Research on character recognition method based on artificial intelligence machine learning[J]. Computer Programming Skills & Maintenance, 2022(9): 154 – 156, 176. (in Chinese)

[4] QU Z G, NIU D Y. Leveraging ResNet and label distribution in advanced intelligent systems for facial expression recognition[J]. Mathematical Biosciences and Engineering, 2023, 20(6): 11101 – 11115.

[5] 曾鹏,李曦,赵璐,等. 基于 MobileNet 和文本识别匹配的证件图片分类算法[J]. 中国人民公安大学学报(自然科学版),2023,29(3):52 – 58.
ZENG P, LI X, ZHAO L, et al. License image classification algorithm based on MobileNet and text recognition matching [J]. Journal of People's Public Security University of China (Science and Technology), 2023, 29(3): 52 – 58. (in Chinese)

[6] DONG C, WANG R F, HANG Y Q. Facial expression recognition based on improved VGG convolutional neural network [J]. Journal of Physics: Conference Series, 2021, 2083(3): 032030.

[7] MA W, GONG C F, XU S B, et al. Multi-scale spatial context-based semantic edge detection[J]. Information Fusion, 2020, 64: 238 – 251.

[8] WANG X F, HE Z H, WANG K, et al. A survey of text detection and recognition algorithms based on deep learning technology[J]. Neurocomputing, 2023, 556: 126702.

[9] 魏永合,宫俊宇. 基于 CNN-LSTM-Attention 的滚动轴承故障诊断[J]. 沈阳理工大学学报,2022,41(4):73 – 77.
WEI Y H, GONG J Y. Fault diagnosis in rolling bearing based on CNN-LSTM-Attention[J]. Journal of Shenyang Ligong University, 2022, 41(4): 73 – 77. (in Chinese)

[10] 华春梦,臧艳辉,马伙财. 一种基于 CRNN 的车牌识别算法研究与应用[J]. 现代信息科技,2021,5(20):78 – 81,86.
HUA C M, ZANG Y H, MA H C. Research and application of a license plate recognition algorithm based on CRNN[J]. Modern Information Technology, 2021, 5(20): 78 – 81, 86. (in Chinese)

[11] 王雪娇,张超敏. 基于 CNN 和 LSTM 的自然场景文本检测应用[J]. 仪表技术,2020(9):17 – 23,45.
WANG X J, ZHANG C M. Application of the natural scene text detection based on CNN and LSTM[J]. Instrumentation Technology, 2020(9): 17 – 23, 45. (in Chinese)

[12] 林金朝,文盼,庞宇. 基于特征金字塔网络的自然场景图像文本检测[J]. 重庆邮电大学学报(自然科学版),2022,34(1):155 – 163.
LIN J Z, WEN P, PANG Y. Text detection of natural scene images based on feature pyramid network[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2022, 34(1): 155 – 163. (in Chinese)

[13] 王文亮,李延祥,张一帆,等. MPANet-YOLOv5:多路径聚合网络复杂海域目标检测[J]. 湖南大学学报(自然科学版),2022,49(10):69 – 76.
WANG W L, LI Y X, ZHANG Y F, et al. MPANet-YOLOv5: multi-path aggregation network for complex sea object detection[J]. Journal of Hunan University(Natural Sciences), 2022, 49(10): 69 – 76. (in Chinese)

[14] TSANG C, BOULTON C, BURGON V, et al. Predicting 30-day mortality after hip fracture surgery: evaluation of the national hip fracture database case-mix adjustment model [J]. Bone & Joint Research, 2017, 6(9): 550 – 556.

[15] AKIKI C, BURGHARDT M. MuSe: the musical sentiment dataset[J]. Journal of Open Humanities Data, 2021, 7: 10.

[16] KONG Q Q, LI B C, CHEN J T, et al. GiantMIDI-Piano: a large-scale MIDI dataset for classical piano music[J]. Transactions of the International Society for Music Information Retrieval, 2022, 5(1): 87 – 98.

[17] LIAO M H, SONG B Y, LONG S B, et al. SynthText3D: synthesizing scene text images from 3D virtual worlds[J]. Science China(Information Sciences), 2020, 63(02): 65 – 78.

[18] PYATAEVA A, GENZA S. Artificial neural network technology for text recognition[J]. CEUR Workshop Proceedings, 2020, 2534: 248 – 252.

(责任编辑:和晓军)