DOI:10.16136/j.joel.2022.05.0761

基于多级特征选择的自然场景文本识别算法

李利荣[1,2*]，张 开[1]，张云良[1]，乐 玲[1]，周 蕾[1]，巩朋成[1,2]

(1.湖北工业大学 电气与电子工程学院，湖北 武汉 430064；2.新能源及电网装备安全监测湖北省工程研究中心，湖北 武汉 430064)

摘要：针对现有场景文本识别方法只关注局部序列字符分类，而忽略了整个单词全局信息的问题，提出了一种多级特征选择的场景文本识别(multilevel feature selection scene text recognition，MFSSTR)算法。该算法使用堆叠块体系结构，利用多级特征选择模块在视觉特征中分别捕获上下文特征和语义特征。在字符预测过程中提出一种新颖的多级注意力选择解码器(multilevel attention selection decoder，MASD)，将视觉特征、上下文特征和语义特征拼接成一个新的特征空间，通过自注意力机制将新的特征空间重新加权，在关注特征序列的内部联系的同时，选择更有价值的特征并参与解码预测，同时在训练过程中引入中间监督，逐渐细化文本预测。实验结果表明，本文算法在多个公共场景文本数据集上识别准确率能达到较高水平，特别是在不规则文本数据集SVTP上准确率能达到87.1%，相比于当前热门算法提升了约2%。

关键词：场景文本识别；特征序列；自注意力机制；多级注意力选择解码器；中间监督

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1005-0086(2022)05-0479-09

Natural scene text recognition algorithm based on multilevel feature selection

LI Lirong[1,2*], ZHANG Kai[1], ZHANG Yunliang[1], YUE Ling[1], ZHOU Lei[1], GONG Pengcheng[1,2]

(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, Hubei 430064, China; 2. Hubei Engineering Research Center for New Energy and Power Grid Equipment Safety Monitoring, Wuhan, Hubei 430064, China)

Abstract: Aiming at the problem that existing scene text recognition methods only focus on the classification of local sequence characters and ignore the global information of the entire word, a multilevel feature selection scene text recognition (MFSSTR) algorithm is proposed. The algorithm uses a stacked block architecture and applies a multilevel feature selection module to capture contextual and semantic features in visual features. In the process of character prediction, a novel multilevel attention selection decoder (MASD) is proposed, which combines visual features, context features and semantic features into a new feature space, and re-weights the new feature space through a self-attention mechanism. While paying attention to the internal relations of the feature sequence, select more valuable features and participate in decoding prediction. At the same time, intermediate supervision is introduced in the training process to gradually refine the text prediction. The experimental results show that the algorithm in this paper can reach a high level of recognition accuracy on multiple public scene text data sets. In particular, the accuracy rate can reach 87.1% on the irregular text data set SVTP, which is improved compared with the current popular algorithms by about 2%.

Key words: scene text recognition; feature sequence; self attention mechanism; multilevel attention se-

* **E-mail**: Rongli@hbut.edu.cn

收稿日期：2021-11-12 **修订日期**：2021-12-14

基金项目：国家自然科学基金(62071172)和新能源及电网装备安全监测湖北省工程研究中心开放研究基金(HBSKF202121)资助项目

lection decoder; intermediate supervision

1 引 言

场景文本识别作为计算机视觉领域中一个研究方向，随着深度学习的发展，在无人驾驶、智能机器人等商业领域中已广泛地应用[1]，但是由于自然场景的环境复杂，导致场景文本识别仍存在很大挑战性。传统的场景文本识别算法通常是逐个字符的识别，但是这类方法有很大的局限性，自然场景中的文本字符难以分割，且传统方法不依赖于上下文之间的依赖关系和字符之间的顺序建模，导致识别效果不理想。

现代的场景文本识别是基于整个单词的识别，将文本图片切分成序列进行预测，避免了每个字符需要单独注释的必要，相比于传统方法能达到更高的准确率。SHI 等[2]提出一种灵活的文本矫正网络，同时提出双向注意力解码结构来捕捉上下文双向的特征信息。SHENG 等[3]提出基于 transformer[4]的场景文本识别模型，并设计了一个模态转换块直接将 2 维的图像转换为 1 维的特征序列。QIAO 等[5]提出一种新的语义模块，并利用预训练的语言模型来监督语义模块来预测语义信息。NEWELL 等[6]提出利用堆叠块体系结构对特征重复地处理，反复获取不同尺度下所包含的信息。BAEK 等[7]提出了一个由 4 个主体步骤构成的 STR(scene text recognition)框架，如图 1 所示，其分为图像转换，特征提取，序列建模和解码预测等 4 个部分。图像转换主要是对输入图像进行空间变化来矫正文本图像，这个过程是可以选择的，但是对不规则的文本图像很重要。特征提取是将提取到的文本特征转化为视觉特征序列的操作，序列中的每一列对应文本图像中不同的感受野，常用的特征提取网络有 ResNet[8]、VGG 等[9]。序列建模是在视觉特征序列中捕获上下文特征信息，然后将两者关联起来用于最后的预测，常用的网络为 LSTM 等[10]。解码预测是输出最终字符序列的过程，常用的方法为 CTC 解码器(connectionist temporal classification decoder)[11]或注意力解码器(attention decoder)[12]，CTC 是将序列中的每一帧分别解码，然后删除空白和重复的字符。注意力解码则更关注于字符间的依赖关系，关注于更重要的信息进行选择性解码。本文以该 STR 主体结构为基线，在它的基础上进行扩展，重新构建一个新的网络模型。

当前场景文本识别方法大多数使用编码器-解码器方法，并且只将视觉特征和上下文特征关联起来，而忽略了全局的语义特征，且极少挖掘多种特征之间的隐藏联系来辅助文本识别。为了解决这一问题，本文提出了一种多级特征选择场景文本识别模型(multilevel feature selection scene text recognition，MFSSTR)，该模型由堆叠块体系构成，其中包括多个由本文设计的多级特征选择模块，每个模块都包括一个双层的 BiLSTM 编码器，通过该编码器在视觉特征中捕获上下文特征，接着将上下文特征通过一个语义模块来预测全局语义特征。在字符预测过程中提出一种新型的多级注意力选择解码器(multilevel attention selection decoder，MASD)，该解码器分为 2 步操作，第 1 步将视觉特征、上下文特征和语义特征这 3 个特征序列组合成一个新的特征空间，通过自注意力机制对特征空间进行加权操作，选择有价值的特征信息的同时抑制其余无关信息，第 2 步将加权后的特征空间通过注意力解码模块来输出最终的字符序列。本文也引入了CTC解码作为中间监督，通过中间监督的方法来优化堆叠块体系结构，同时提升模型的效率。

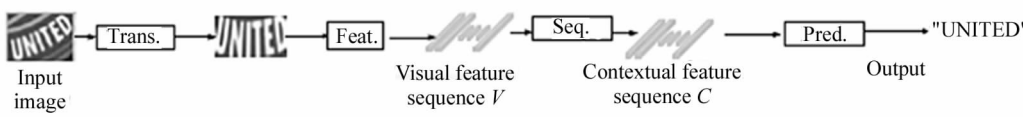

图 1　Baek 等提出的场景文本识别四步框架

Fig. 1　The four-step framework for scene text recognition proposed by Baek et al.

2 算法框架

本文所设计的 MFSSTR 模型如图 2 所示，首先利用薄板样条插值(thin plate spline，TPS)[13]变换对输入图像进行矫正，这个过程可以提高模型识别的效率和准确率。其次使用了一种改进的 39 层残差网络(ResNet-39)在文本图像中提取视觉特征，使字符沿着水平方向切分为特征序列，同时仍保留最佳的表征能力和感受野信息。接着提出基于堆叠块体系结构的多级特征选择模块，每个模块都通过 BiLSTM编码和语义模块在视觉特征中获取上下文

信息和语义信息,然后将视觉特征、上下文特征和语义特征这 3 个不同层面的特征信息拼接成一个新的特征空间,将得到的新特征空间通过 MASD 模块进行解码操作,该解码过程分为 2 步,第 1 步是在特征空间中选择重要的特征信息,第2步是通过注意力解码关注特征序列的内部关系,通过特征内部之间的信息交互来预测输出序列。本文在模型训练的过程中引入了中间监督的方法,用于更好地训练深层结构的堆叠块体系,同时提升了 MASD 解码的性能。

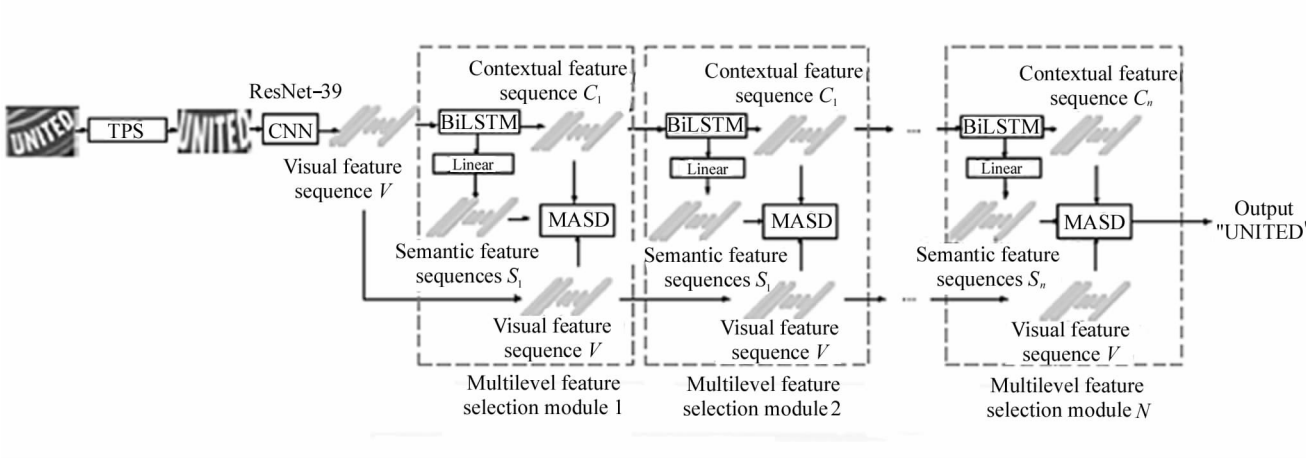

图 2　多级特征选择场景文本识别(MFSSTR)模型

Fig. 2　Multilevel feature selection scene text recognition module

2.1　文本矫正模块

自然场景中不规则的图片文本较多,整齐排列和分辨度高的文字信息对后续文本识别尤为重要,因此本文使用 TPS 方法进行输入图片的文本矫正操作,TPS 矫正模块如图 3 所示,TPS 是在一组基准点之间采用平滑样条插值操作,具体而言,它通过检测文本区域顶部和底部的基准点,并控制这些基准点包围的区域得到矫正后的文本图像,输出图像大小固定为32×100,作为后续特征模块的输入大小。

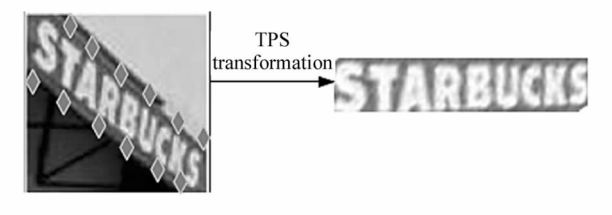

图 3　TPS 矫正模块

Fig. 3　TPS rectification module

2.2　特征提取模块

本文使用了一种改进的 ResNet-39 对矫正后的文本图像进行特征提取,其结构如表 1 所示,第 1 层使用了一个的卷积将输入图片维度变为 32,后面每个残差块都由与的卷积核构成,且每个阶段的残差块数量分别为 3、4、6、3、3,在第 2 和第 3 阶段使用步长为的卷积核对特征进行下采样操作,后面第 4、5、6 阶段则将步长进行调整,这样做的目的是在下采样的过程中宽度上保留合适的特征信息,同时使输出的特征图的高度为 1,可以更好的让特征图沿水平方向切分为特征序列。

表 1　ResNet-39 网络结构

Tab. 1　ResNet-39 network structure

Stage	Layers	Configurations	Output
1	Conv	3×3conv,1×1stride	32×100
2	Block1	$\begin{bmatrix}1\times1\text{conv},32\\3\times3\text{conv},32\end{bmatrix}\times3$,2×2stride	16×50
3	Block2	$\begin{bmatrix}1\times1\text{conv},64\\3\times3\text{conv},64\end{bmatrix}\times3$,2×2stride	8×25
4	Block3	$\begin{bmatrix}1\times1\text{conv},128\\3\times3\text{conv},128\end{bmatrix}\times3$,2×2stride	4×25
5	Block4	$\begin{bmatrix}1\times1\text{conv},256\\3\times3\text{conv},256\end{bmatrix}\times3$,2×1stride	2×25
6	Block5	$\begin{bmatrix}1\times1\text{conv},512\\3\times3\text{conv},512\end{bmatrix}\times3$,2×1stride	1×25

通过本文改进的网络结构最终输出大小为 1×25,通道数为 512 的特征图。假设特征提取模块能够获得大小为 $D\times H\times W$ 的更细化的文本特征图,其中 H,W 和 D 分别代表输出文本特征图的高度、宽度和通道数,然后将特征图进行 Map-to-Sequence[11]操作,特征图转换为特征序列如图 4 所示,沿着横轴方向将特征图转化成 W 个视觉特征序列,

则每个特征序列的维度为 $H\times D$，视觉特征序列可以表示为 $V(v_1,v_2,\cdots,v_n)$，其中 n 等于特征图的宽度 W。

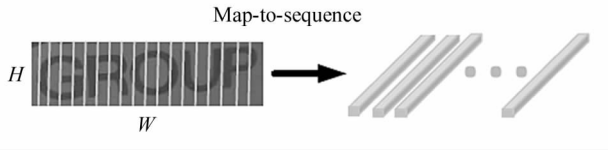

图 4　特征图转换为特征序列

Fig. 4　Transform feature map into feature sequence

2.3　多级特征选择模块

场景本文通常包含多个层次的信息特征，其中视觉特征包含丰富的图像特征信息，但是视觉特征往往因为感受野的大小而受到局限，而为了获取文本的上下文特征信息，当前大多数方法对视觉特征进行序列建模获取上下文特征，然而有限的上下文特征不能处理没有全局信息的字符，因此全局的语义特征对最后结果的准确预测尤为重要。本文设计的多级特征选择模块(multilevel feature selection module)见图 2 中虚线框部分。本文将视觉特征序列 V 输入到双层的 BiLSTM 网络中得到上下文特征序列 C，表示为 $C=(c_1,c_2,\cdots,c_n)$，BiLSTM 网络由双向的 LSTM 组成，它可以同时从 2 个方向捕获上下文信息特征，从而更好的关注到输入特征不同通道之间的潜在联系，更完整的学习文本信息特征，BiLSTM 编码器如图 5 所示，它将特征序列通过上下 2 个方向输入到其中，每个 t 时刻的输入视觉特征 v_t 分别计算出对应的隐藏状态向量 L_t，再通过一系列计算得到每个时刻的输出上下文特征 c_t，其表达式为：

$$c_t=\sigma(W(c_{t-1},v_t)+b)\times\tanh(L_t),\tag{1}$$

式中，W 和 b 分别表示训练权重和偏置参数，σ 表示 Sigmoid 激活函数，c_{t-1} 表示在时间 t 上一个时刻的上下文特征，L_t 是 t 时刻的每个 LSTM 单元隐藏状态。

为了挖掘全局语义辅助文本识别，本文通过一个语义模块在上下文特征中获得语义特征序列 S，具体而言该模块可以用线性函数来描述其过程(见图 2 中 Linear 模块)。首先将上下文特征序列 C 展开为 1 维的特征向量 $\boldsymbol{I}$，其维度为 K，具体的 K 等于 $H\times D$，H 和 D 分别为序列所对应的特征图的高度和通道数，然后通过线性函数将其转化为语义特征序列 S，其表达式为：

$$S=W_2\sigma(W_1\boldsymbol{I}+b_1)+b_2,\tag{2}$$

式中，W_1 和 W_2 为训练权重，b_1 和 b_2 为偏置参数，σ 为 ReLu 激活函数。

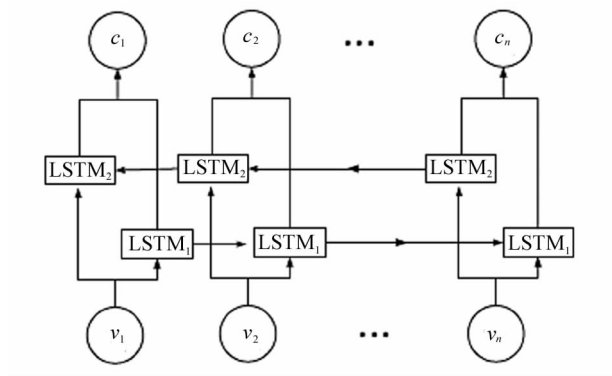

图 5　BiLSTM 编码器

Fig. 5　BiLSTM encoder

本文将视觉特征、上下文特征和语义特征三者结合起来生成一个新的特征空间 $M=(V,C,S)$，特征空间 M 既用于多级注意力选择解码如图 6 所示，也用作于下一个多级特征选择模块的输入，如图 7 所示，视觉特征 V 在每一个多级特征选择模块中保持不变，上下文特征 C 和语义特征 S 则逐级更新，第 n 个模块的表示为 $M_n=(V,C_n,S_n)$，第 $n+1$ 个模块使用 C_n 作为双层 BiLSTM 的输入，得到输出 C_{n+1}，再将 C_{n+1} 输入到语义模块中得到 S_{n+1}，从而第 $n+1$ 个特征空间变为 $M_{n+1}=(V,C_{n+1},S_{n+1})$。这些多级特征选择模块可以根据任务需要多次堆叠在一起，视觉特征 V 在多级特征选择模块中没有更新，最终的预测由最后一个多级特征选择模块中的解码器提供。

2.3.1　多级注意力选择解码器

MASD 具体包含 2 个步骤，如图 7 所示，第 1 步本文使用自注意力(self attention)机制[4]对特征空间 M 进行处理，自注意力关注于特征空间中重要的特征信息，通过一个全连接层产生该特征空间的注意力图，然后将注意力图对原特征空间 M 相乘，得到注意力加权过后的特征空间 M'。不同于传统的 RNN 序列模型，自注意力机制的输出是兼顾了所有输入的全局信息，而且它的输出是同时并行化计算得到的，并不需要再按照依次顺序得到，极大的提升了效率。自注意力结构有三个分支，如图 6 所示，分别为 query($\boldsymbol{Q}$)、key($\boldsymbol{K}$)、value($\boldsymbol{V}$)，首先将 query 和对应的 key 进行相似度计算得到权重，其次使用 softmax 函数对权重进行归一化处理，将权重和相应的 value 进行加权求和得到最终输出，其表达式为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{soft max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right), \quad (3)$$

式中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 3 个矩阵分别由输入序列进行一系列线性计算得到,$\sqrt{d_k}$ 为矩阵 $\boldsymbol{K}$ 的维度。

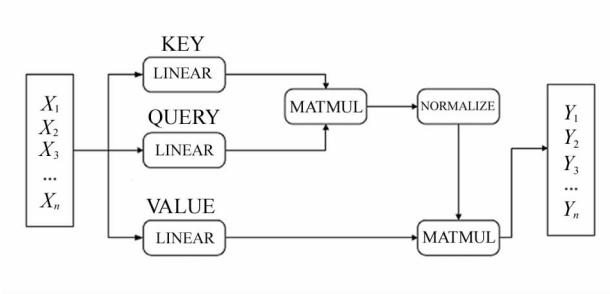

图 6　自注意力机制的基本结构

Fig. 6　The basic structure of the self-attention mechanism

第 2 步是将加权后得到的特征空间 M' 当作一个序列并提取它的内部之间的关系,M' 的解码是通过单独的注意力解码器完成的,对于每个时间步 t,通过解码器输出 y_t,解码的整个流程从计算注意力权重 α 开始,其表达式为:

$$e_{t,i} = \omega^{T}\tanh(W_{s_{t-1}} + Vm'_i + b), \quad (4)$$

$$\alpha_{t,i} = \exp(e_{t,i}) / \sum_{i=1}^{n} e_{t,i^*}, \quad (5)$$

式中,b、ω、V、W 都是训练参数,s_{t-1} 是解码器中循环单元在时间 t 前一步的隐藏状态,m' 表示 M' 中的一列,解码器 M' 将中每个元素使用线性方式组合为向量 $\boldsymbol{G}(g_1, g_2, \cdots, g_t)$,本文称为注意力向量,其表达式为:

$$g_t = \sum_{i=1}^{n} \alpha_{t,i} m'_i。 \quad (6)$$

由式(6)可知,注意力向量 g_t 表示了用 M' 编码的整个空间特征的一小部分,将这一小部分送入到后续解码器单元中,得到输出向量 x_t 和状态向量 s_t,其表示为:

$$(x_t, s_t) = RNN(s_{t-1}, (g_t, f(y_{t-1}))), \quad (7)$$

$$p(y_t) = \text{softmax}(W_o x_t + b_o), \quad (8)$$

式中,$(g_t, f(y_{t-1}))$ 表示为注意力向量 g_t 和 y_{t-1} 运算后的拼接形式。$RNN(\cdot)$ 函数为循环神经网络单元的函数,通过该函数的输出 x_t 送入到式(8)中来预测给定字符 y_t 的概率,式(8)中 W_o 和 b_o 为训练参数,softmax 为归一化指数函数,得到解码之后的输出序列 $Y(y_1, y_2, \cdots, y_n)$,最终实现每一个字符的预测。

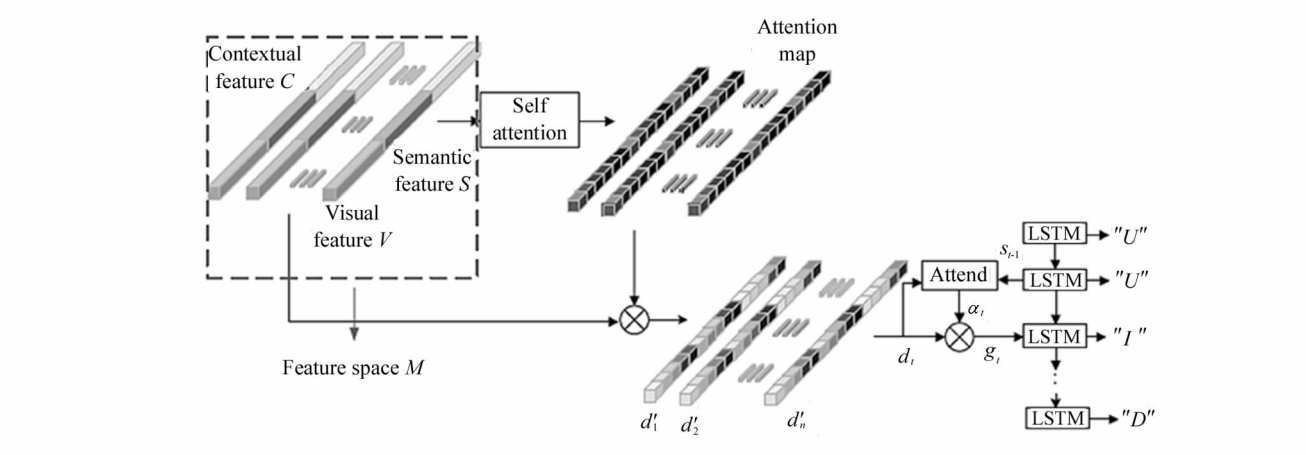

图 7　多级注意力选择解码

Fig. 7　Multilevel attention selection decoding

2.4　损失函数

本文的损失函数由 2 部分组成,为由多级注意力选择解码和提供中间监督的 CTC Loss 构成,其表达式为:

$$L = \sum_{i=1} \lambda_i L_{\text{att},i} + \lambda_{\text{CTC}} L_{\text{CTC}}, \quad (9)$$

式中,$\sum_{i=1} \lambda_i L_{\text{att},i}$ 表示所有多级注意力选择解码的损失之和,L_{CTC} 表示基于CTC解码的损失函数,CTC解码器是将视觉特征 V 进行中间解码,细化每个特征的表达形式并将它们分类为单独的符号,对最终的解码过程提供中间监督并完善预测结果,并对本文训练堆叠性的深层网络提供良好的帮助。其中 λ 是作为超参数来平衡不同的监督过程,λ_{CTC} 和 λ_i 分别设置为 0.1 和 1,基于注意力机制的损失函数[14] 表示为:

$$L_{\text{att}} = -\sum_{t=1} \log P(y_t \mid I, \theta) \quad (10)$$

式中，y_t 表示第 t 个特征的真实值，I 为参与解码过程的特征序列，θ 为所有网络参数组合成的向量。

基于 CTC 中间监督的损失函数表示为：

$$L_{CTC}=-\sum_{l_i\in x}\log P(l_i\mid y_i), \tag{11}$$

式中，x 表示全部训练样本，l_i 代表真实标签的序列，y_i 代表解码之后预测得到的序列。

3 实验分析

3.1 数据集

本次实验所采用的训练集为 Synth90k[15] 和 ICDAR2015[16] 数据集，Synth90k 是人工合成的大型场景文本训练数据集，包含上百万张图片，每个图片中的单词均有真实框裁剪并注释。ICDAR2015 则是常用的文本识别比赛的数据集，通过谷歌眼镜收集自然街景得到，包括 2077 张裁剪后的单词图像。测试集选择 6 个公共数据集，均为裁剪好的英文单词及数字图像并带有标签。测试集的相关描述如下所示：

1) 规则文本数据集

ICDAR2013[17](IC13)数据集以街道广告牌和路标等清晰拍摄照片为主，包含 1015 张裁剪单词的图像。IIIT5K[18] 是在网络上随机搜索的街景图片组成，由 3000 张裁剪好的文本图片构成。SVT[19] 来自于谷歌街景图像库，大部分图片较为模糊且清晰度不高，但是拍摄角度对焦，包含 647 张裁剪后的场景文本单词图片。

2) 不规则文本数据集

SVTP[20] 来自于谷歌自然街景拍摄图像库，所有图片为非正面视角拍摄，包含很多变形的文本图像，总共有 639 张裁剪后的单词图像。CUTE80[21] 数据集一共是 80 张来源自然场景中拍摄的高分辨率图像，其中包含大量的弯曲文本，后裁剪为 288 张单词图片用作测试集。

3.2 评价指标

本文使用文本识别全对准确率，完全正确识别单词中每一个字符作为测试阶段的评价指标，其表达式为：

$$Accuracy=\frac{C}{N}, \tag{12}$$

式中，C 表示为完全识别的样本数量，N 表示测试集的全部样本。

3.3 实验细节

首先将 Synth90k 数据集在本文提出的网络上训练 2 个 epoch 作为预训练模型，然后利用 ICDAR2015 数据集中全部的数据样本训练 600 个 epoch 对模型进行微调，Batch_size 设置为 64，使用 Adam 优化器在训练过程中对模型参数进行更新，初始动量设置为 0.9，权重衰减系数设置为 0.005，训练过程中总的损失函数曲线图如图 8 所示。

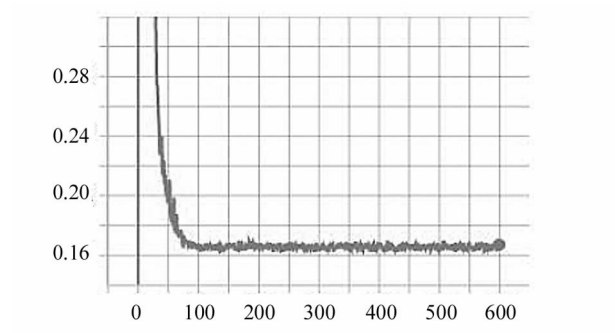

图 8 总的损失函数曲线图

Fig. 8 The total loss function curve

3.4 消融实验

为了更好的验证本文算法的优越性，实验以文献提出的模型框架为基线(Baseline)，在 6 个数据集上进行消融实验对比，在分别列出每个数据集的测试结果的基础上，本文将 4 个规则数据集(IIIT5K，SVT，IC03，IC13)和 2 个不规则的数据集(SVTP，CUTE)分为 2 大类，分别根据样本数量得到加权后的准确率平均值。对比注意力解码器(Attention decoder)和本文提出的 MASD 的效果，同时对比引入 CTC 解码作为中间监督对模型性能的影响。

规则与不规则文本数据集消融实验如表 2 和 3 所示。从表 2 和表 3 中的数据可以得到，添加了一个用于中间监督的 CTC 解码器后，基线模型在规则文本和不规则文本上的准确率分别提升了0.3%和1.4%，本文提出的MFSSTR模型则分别提高了0.3%和0.4%，证明了中间监督的有效性，能提升最终解码的效果。而在使用本文提出的 MASD 替换掉基线模型中的注意力解码器后，在同样使用 CTC 作为中间监督时，本文的 MFSSTR 模型比基线模型分别上升了 0.6%和 2.9%，表明了 MASD 比注意力解码器有更好的优越性。

图 9 是常规的注意力解码器和本文提出的 MASD 在场景文本中的热力图可视化对比，可以得出 MASD 相比于注意力解码器能够更完整的覆盖每个字符区域，且充分利用到相互关联的文字信息，证明了该方法的有效性。

从表 4 可以分析出,在同样使用中间监督的情况下,通过在实验中增加多级特征选择模块数量,本文的模型在各个数据集上的性能逐渐提升,在模块数量达到5个时整体表现最优,说明了堆叠块体系结构进行重复特征处理可以保证稳定且优良的训练效果。而超过 5 个后整体略微下降,也说明模块数量增加后导致更深层的模型结构出现退化问题,如何改进训练方法或提升训练技巧,也是本文未来工

表 2　规则文本数据集消融实验

Tab. 2　Ablation experiment of regular text data set

Method	CTC decoder	Attention decoder	MASD	*N* modules	IIIT5K	SVT	IC13	Regular text average
Baseline	—	√	—	—	93.1	88.5	92.8	92.4
Baseline	√	√	—	—	92.7	90.1	93.5	92.5
MFSSTR	—	—	√	1	93.3	89.4	93.8	92.8
MFSSTR	√	—	√	1	**93.6**	**89.2**	**93.9**	**93.1**

注:*N* modules 表示多级特征选择模块的数量,加粗部分代表最优值。

表 3　不规则文本数据集消融实验

Tab. 3　Ablation experiment of irregular text data set

Method	CTC decoder	Attention decoder	MASD	*N* modules	SVTP	CUTE	Irregular text average
Baseline	—	√	—	—	81.4	80.1	81.0
Baseline	√	√	—	—	83.3	80.3	82.4
MFSSTR	—	—	√	1	84.2	**86.5**	84.9
MFSSTR	√	—	√	1	**84.9**	86.1	**85.3**

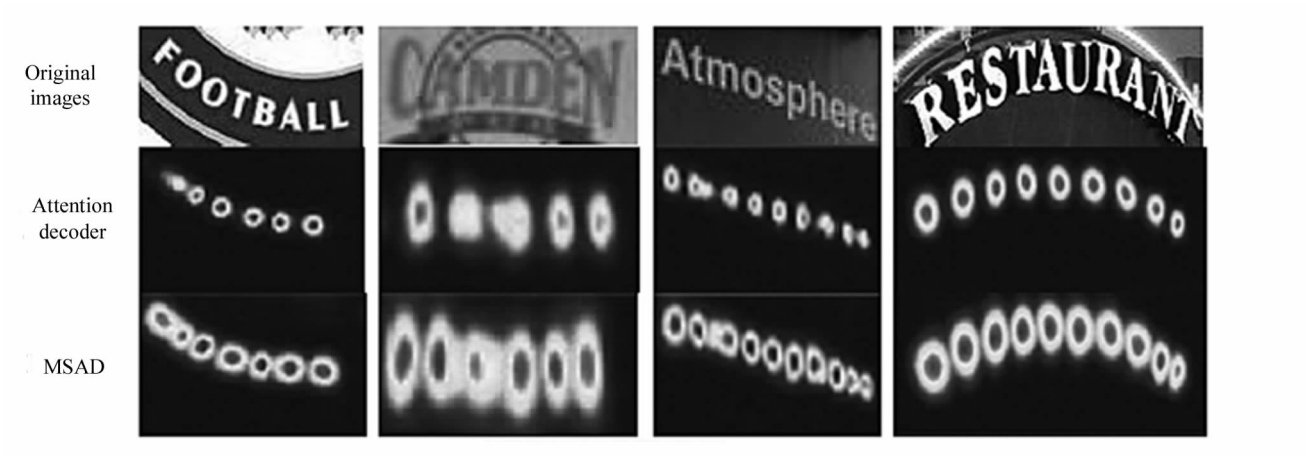

图 9　热力图可视化对比

Fig. 9　Visual comparison of heat map

表 4　多级特征选择模块数量对比实验

Tab. 4　Comparison experiment of the number of multilevel feature selection module

Method	*N* modules	IIIT5K	SVT	IC13	SVTP	CUTE	Regular text average	Irregular text average
MFSSTR	1	93.6	89.2	93.9	84.9	86.1	93.1	85.3
MFSSTR	2	93.7	90.7	94.5	86.1	86.8	93.5	86.3
MFSSTR	3	93.9	90.4	94.7	86.5	86.4	93.6	86.5
MFSSTR	4	94.1	91.2	94.8	**87.3**	86.5	93.9	87.1
MFSSTR	5	**94.6**	**92.8**	**95.7**	87.1	**87.6**	**94.6**	**87.3**
MFSSTR	6	93.3	91.0	94.2	86.3	84.9	93.2	85.9

作需要解决的问题。

3.5 对比实验

表5为本文算法与近几年的主流算法的对比实验，本文算法在规则文本数据集SVT和IC13上准确率分别能达到92.8%和95.7%，在不规则文本数据集SVTP上达到87.1%，均优于2020年提出SEED算法和SRN算法。在IIIT5K上和CUTE上准确率分别达到94.6%和87.6%，仅略低于Mask TextSpotter算法，在实验中使用的3个规则文本数据集上的平均准确率能达到94.6%，整体高于平均水平且具有很强竞争性。在另外2个不规则文本数据集上能达到87.3%，相比于其它算法高出至少2%，证明了本文算法的有效性。

表5 本文MFSSTR模型与当前主流文本识别模型对比实验

Tab. 5 Comparative experiment between the MFSSTR model in this paper and the current mainstream text recognition model

Method	IIIT5K	SVT	IC13	SVTP	CUTE	Regular text average	Irregular text average
CRNN[11]	78.3	80.9	86.8	—	—	80.5	—
FAN[14]	87.3	85.9	93.2	—	—	88.4	—
ASTER[2]	93.5	89.4	91.9	78.5	79.4	92.6	78.8
MORAN[22]	91.1	88.4	92.4	76.2	77.5	91.0	76.6
Mask TextSpotter[23]	**95.2**	91.6	94.9	83.7	**88.2**	94.6	85.1
ScRN[24]	94.3	88.7	93.8	80.8	87.3	93.4	82.8
SEED[5]	93.8	89.6	92.8	81.4	83.6	93.0	82.1
SRN[25]	94.8	91.5	95.5	85.1	87.8	94.4	85.9
MFSSTR	94.6	**92.8**	**95.7**	**87.1**	87.6	**94.6**	**87.3**

4 结　论

本文提出了一种MFSSTR算法，该方法利用堆叠块的多级特征选择模块构建一个深层的网络模型，通过该模型在视觉特征中捕获上下文特征和全局语义特征，并将三者特征结合起来构建一个新的特征空间，同时随着模型深度的增加，获取的特征信息也逐级更新，本文还在训练过程中引入中间监督来细化特征并完善预测结果。在字符预测阶段本文提出一种新颖的MASD，关注特征序列之间的内部联系，选择重要的信息特征参与预测解码。实验结果表明，本文提出的算法在多种场景公共数据集上均取得很好的效果，证明了该算法的有效性。在未来工作中，轻量化的端到端场景文本检测与识别算法将成为研究的重点。

参考文献：

[1] LIU C Y, CHEN X X, LUO C J, et al. Deep learning method for natural scene text detection and recognition[J]. Journal of Image and Graphics, 2021, 26(6): 1330-1367.
刘崇宇，陈晓雪，罗灿杰，等. 自然场景文本检测与识别的深度学习方法[J]. 中国图象图形学报，2021，26(6)：1330-1367.

[2] SHI B, YANG M, WANG X, et al. ASTER: an attentional scene text recognizer with flexible rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2035-2048.

[3] SHENG F, CHEN Z, XU B. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition [C]//International Conference on Document Analysis and Recognition, September 20-25, 2019, Sydney, NSW, Australia. New York: IEEE, 2019: 781-786.

[4] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada. New York: IEEE, 2021: 21-25.

[5] QIAO Z, ZHOU Y, YANG D, et al. Seed: semantics enhanced encoder-decoder framework for scene text recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 13525-13534.

[6] NEWELL A, HUANG Z, DENG J. Associative embedding: end-to-end learning for joint detection and grouping[EB/OL]. (2017-06-09) [2021-11-12]. https://arxiv.org/abs/1611.05424.

[7] BAEK J, KIM G, LEE J, et al. What is wrong with scene text recognition model comparisons dataset and model

analysis[C]//IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE, 2019: 4714-4722.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-08-10) [2021-11-12]. https://arxiv.org/abs/1409.1556.

[10] GRAVES A, LIWICKI M, FERNANDEZ S, et al. A novel connectionist system for unconstrained handwriting recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 855-868.

[11] 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 Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298-2304.

[12] LITMAN R, ANSCHEL O, TSIPER S, et al. SCATTER: Selective context attentional scene text recognizer [EB/OL]. (2020-03-25) [2021-11-12]. https://arxiv.org/abs/2003.11288.

[13] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4168-4176.

[14] CHENG Z, BAI F, XU Y, et al. Focusing attention: towards accurate text recognition in natural images [C]//IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5086-5094.

[15] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Synthetic data and artificial neural networks for natural scene text recognition[EB/OL]. (2014-12-09) [2021-11-12]. https://arxiv.org/abs/1406.2227v4.

[16] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]//International Conference on Document Analysis and Recognition, August 23-26, 2015, Nancy, France. New York: IEEE, 2015: 1156-1160.

[17] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]//International Conference on Document Analysis and Recognition, August 25-28, 2013, Washington, DC, USA. New York: IEEE, 2013: 1484-1493.

[18] MISHRA A, ALAHARI K, JAWAHAR C V. Scene text recognition using higher order language priors[C]//British Machine Vision Conference, September 3-7, 2012, Guildford, Surrey, UK. Durham: British Machine Vision Association (BMVA), 2012: 1-11.

[19] WANG K, BABENKO B, BELONGIE S. End-to-end scene text recognition[C]//IEEE International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE, 2011: 1457-1464.

[20] PHAN T Q, SHIVAKUMARA P, TIAN S, et al. Recognizing text with perspective distortion in natural scenes[C]//IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE, 2013: 569-576.

[21] RISNUMAWAN A, SHIVAKUMARA P, CHAN C S, et al. A robust arbitrary text detection system for natural scene images[J]. Expert Systems with Applications, 2014, 41(18): 8027-8048.

[22] LUO C, JIN L, SUN Z. MORAN: A multi-object rectified attention network for scene text recognition[J]. Pattern Recognition, 2019, 90: 109-118.

[23] LYU P, LIAO M, YAO C, et al. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[EB/OL]. (2018-08-01) [2021-11-12]. https://arxiv.org/abs/1807.02242v2.

[24] YANG M, GUAN Y, LIAO M, et al. Symmetry-constrained rectification network for scene text recognition[C]//IEEE International Conference on Computer Vision, October 27-28, 2019, Seoul, Korea (South). New York: IEEE, 2019: 9146-9155.

[25] YU D, LI X, ZHANG C, et al. Towards accurate scene text recognition with semantic reasoning networks[EB/OL]. (2020-03-27) [2021-11-12]. https://arxiv.org/abs/2003.12294v1.

作者简介:

李利荣　(1974—),女,博士,讲师,硕士生导师,主要从事计算机视觉与深度学习方面的研究.