文章编号：2096-1618(2023)06-0649-07

基于注意力多分支卷积和 Transformer 的手写文本识别

郑晓旭， 舒珊珊， 文成玉
(成都信息工程大学通信工程学院，四川 成都 610225)

摘要：手写体识别技术作为自动阅卷的关键一环受到广泛研究。针对中文手写文本字迹复杂的问题，提出一种文本定位和识别的手写汉字文本识别方法。在文本定位信息中使用透视变化纠正倾斜的文本，特征提取阶段使用注意力多分支卷积层提取文本图像关键区域特征以及多尺度特征融合，语义提取阶段通过时间卷积网络和 Transformer 编码器构建序列信息和建模上下文语义信息，最后以链接时序分类函数，实现序列特征和字符序列标签对齐。所提方法在公开数据集 CASIA-HWDB 上进行实验，结果表明，注意力分支卷积层和语义提取层有效提升算法性能，证明所提方法的可行性。

关 键 词：手写文本识别；Transformer；注意力机制；链接时序分类
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2023.06.005

0 引言

智能阅卷、银行支票识别、自动入档等解放人类劳动力的应用需求，催促着自动文字识别技术的发展。汉字字符种类繁多，不同字符间相似度高，书写者书写风格复杂多变，字迹下倾上漂等对识别算法造成极大干扰，使其成为模式识别领域中的热点研究问题。

基于分割的手写文本识别，将输入的文本图片进行分割，获得整个或部分字符的一系列片段，组合这些片段生成候选项，利用字符分类器和上下文信息完成识别[1-2]。由于粘连字符难以切分，错误切分对连续识别造成影响，额外的后处理过程导致模型设计过于复杂，识别算法开始向无分割的方法衍生。无分割的方法主要利用滑动窗口滑动步长，通过分类器对窗口内字符进行识别。识别过程无需任何字符或单词切分，避免字符切分错误对识别精度的影响。如 Su 等[3]提出高斯混合隐马尔可夫模型(hidden markov model，HMM)对滑动窗口内的字符进行识别。

深度卷积神经网络赢得图像分类挑战后，基于深度神经网络的无分割手写识别方法不断被提出。从模型的结构和切入角度分析，可分为面向文本行和文本页识别的模型。文本行识别模型采用编-解码或特征对齐等技术，将输入的文本行图像看作多字符序列映射问题。Shi 等[4]结合卷积神经网络和循环神经网络，提出卷积循环神经网络模型(convolutional recurrent neural network，CRNN)直接运行于单词标签上，CNN 进行特征提取，RNN 建模序列信息。Messian 等[5]利用多维长短时记忆(long-short term memory，LSTM)循环神经网络进行端到端文本识别。上述模型表现出良好性能，但存在以下限制：循环网络对于长序列文本联系的利用并不充分，其序列信息生成依赖于循环迭代过程，造成强烈耦合；固定感受野的 CNN 模型，对于脱机汉字大小不一致的特点，提取的特征表达力不够，导致泛化性弱。

文本页图片无分割的识别方法[6-7]，通过拉伸、挤压方式将整个文本页图片逐渐压缩成几行或一整行特征图进行识别。该策略丢失文本的定位信息，无法处理倾斜文本，复杂的层次结构会加重识别难度，识别性能有待提高。

为解决上述问题，本文提出一种基于注意力多分支卷积和 Transformer 的手写文本识别算法。通过文本纠正模块进行倾斜计算纠正文本，由注意力多分支卷积实现变感受野和注意力机制结合，从而聚焦文本图像重要特征，通过自注意力机制捕获序列长距离语义关系。

1 算法框架

1.1 总体框架

本文所提出的算法框架主要由检测网络和识别网络两部分构成，如图 1 所示。检测网络主要完成文本定位和倾斜信息获取，识别网络主要完成文本纠正、特征提取和文本预测识别。

收稿日期：2022-12-09

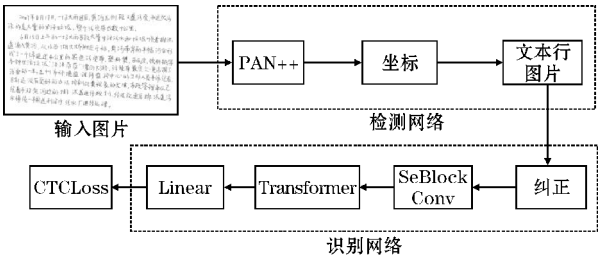

图 1 网络框架

其中,识别网络包含文本纠正模块、注意力多分支提取网络、语义提取层和输出层,具体设置如下:

(1)检测网络用于将输入文本图片转换为单文本行表示形式,使用包含位置信息的坐标点表示不同行文本所在区域,同时坐标点暗含了文本的倾斜信息。该层的输入为整张文本,输出是文本信息坐标点。

(2)识别网络中文本纠正模块用于纠正倾斜文本,采用透视变换对文本图片进行空间映射变换。该层输入的是文本行图片和位置信息坐标点,输出的是纠正后图片。

(3)在注意力多分支卷积层中,分别对每层的特征图进行基于空间和通道注意力的细化特征提取,以及变化感受野堆叠的分支特征融合,最后得到字符特征表达形式。该层输入的是文本图片,输出提取的文本字符特征。

(4)语义提取层将上一步的字符特征转变为包含上下文的序列信息特征,使用 TCN 做序列特征提取,Transformer 使用自注意力机制融合文本上下文语义信息,输出的是基于自注意力权重的序列特征。

(5)输出层通过链接时序分类 CTC 做序列特征对齐,实现表征序列到文本序列的转化,完成文本预测识别。该层输入是语义提取层获取的时间步特征,输出是整张文本的识别文本。

1.2 检测网络

在检测网络产生文本行定位信息,选用 PAN++[8] 检测网络作为文本定位模型。模型采用语义分割的方法,能检测任意形状的文本。每行文本视为周围像素包裹的文本中心核,不同文本核之间存在间隔以此区分不同文本行。选择 ResNet[9] 作为骨架网络,块堆叠数目设置为 3、3、9、3,滑动步长设置为 2,在每个残差堆叠块中引入深度可分离卷积来减少网络参数量,沿用其特征增强模块(FPEMv2),以融合不同尺度的特征信息。检测效果如图 2 所示。

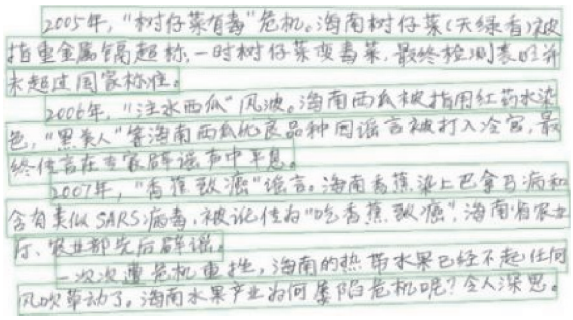

图 2 文本检测结果图

1.3 识别网络

1.3.1 文本图片纠正

文本检测网络只涉及文本的定位,在无约束的条件下,手写文本上漂下倾,对识别造成影响,识别前利用纠正算法对文本进行水平纠正。检测网络的定位信息包含 4 个顶点坐标,采用透视变换纠正倾斜的文本行。透视变换把图片投影到一个新的视平面,从二维平面转换到三维空间,再映射到另一个二维平面。变换矩阵由给定的 4 个顶点坐标和目标坐标计算可得,变换公式:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

式中,(x,y) 为原始图片的坐标点,变换的目标坐标为 (x',y'),展开可得:

$$\begin{cases} x' = \dfrac{X}{Y} = \dfrac{a_{11}x + a_{12}y + a_{13}}{a_{31}x + a_{32}y + a_{33}} \\ y' = \dfrac{Y}{Z} = \dfrac{a_{21}x + a_{22}y + a_{23}}{a_{31}x + a_{32}y + a_{33}} \end{cases} \tag{2}$$

由文本行顶点信息可获得变换后目标坐标点。首先进行倾斜计算:

$$\theta = \arccos\left(\frac{\sqrt{(x_1 - x_4)^2 + (y_1 - y_4)^2}}{x_4 - x_1} \times \frac{180}{\pi} \right)$$
$$\theta = -\theta, y_4 > y_1 \tag{3}$$

计算倾斜角 θ 表示倾斜程度,正负表示文本上倾或下斜,(x_1,y_1)、(x_4,y_4) 表示左上、右上顶点。

然后,根据得到的角度 θ 变换至水平位置的目标坐标点 (x',y')。变换前后的 8 个坐标点利用式(2)得到变换矩阵,通过矩阵透视变换文本行至水平方向,获得纠正后的文本图片,如图 3 所示。

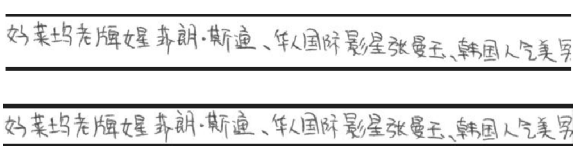

图 3 文本行纠正结果图

1.3.2　识别框架

文本图像转换为字符序列对齐,需要获得细粒度的字符特征,要求特征提取网络能突出文本图像的重要区域,有效提取各个字符。从图像的全局特征角度设计含有注意力机制的多分支卷积层,关注特征图中重要区域,以及增强重要特征通道。语义特征层则实现字符特征到序列特征映射,建模特征序列上下文信息以构建序列特征间的语义联系。图 4 为识别网络总体结构,其中虚线框表示特征提取操作,主要由注意力多分支卷积和下采样来实现,实线框表示语义提取操作,主要由 TCN 和 Transformer 来实现,后通过链接时序分类完成文本识别。

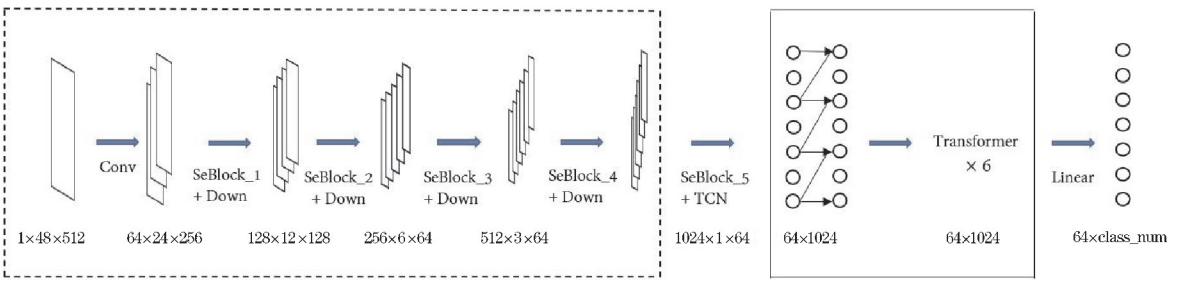

图 4　识别网络结构图

通过堆叠注意力多分支卷积和下采样层进行输入图片的高维特征提取,注意力分支卷积层中包含不同数量的多分支卷积和多层感知机,构成 SeBlock_1 ~ SeBlock_5 五个特征提取层,每层后添加下采样层,以此缩减输入图片尺寸和通道增加,由 $h \times w$ 减小至 $1 \times w/8$,通道数由 1 增加至 1024。

语义提取层,通过时间卷积网络引导字符特征到序列特征映射,主要由 4 层因果卷积层和空洞卷积所构成,每层空洞数翻倍;通过 Transformer 建模特征序列上下文信息,主要由 6 层堆叠的自注意力编码器构成,该层维持特征图大小和通道数不变。另外,在最后一层 Transformer 后连接线性层,通道数由 1024 变为分类类别数。

1.3.3　注意力多分支卷积层

由于不规范的书写会严重影响网络的判别力,如连笔拖拽、部首分离,使得卷积网络特征提取过程关注干扰区域,造成误判。注意力机制[10]模拟了人眼的视觉感知,通过对不同区域进行注意力映射,以减小干扰信息的权重输入,聚焦有用信息的提取。在卷积网络中引入空间和通道注意力,利用空间定位和通道压缩,实现跨通道和空间信息整合,提升网络的关键信息提取能力和过滤背景噪声。经过多层卷积后,原始高维特征图亦含有重要语义信息,在空间和通道注意力残差连接输入特征,利用原始输入特征提升深度神经网络收敛性,空间和通道注意力结构如图 5 和图 6 所示。

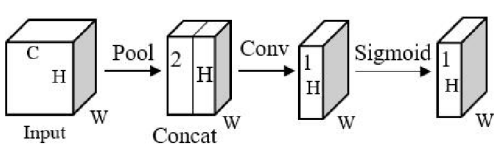

图 5　空间注意力结构图

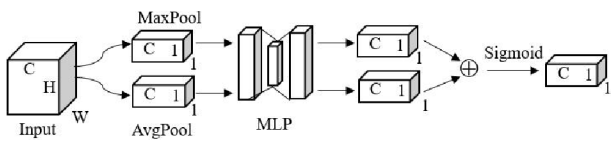

图 6　通道注意力结构图

空间注意力使用最大池化和平均池化进行输入特征 $I_n \in R^{N\times C\times H\times W}$ 通道维度压缩,得到输入纹理特征信息 $F_m \in R^{N\times 1\times H\times W}$ 和背景特征 $F_v \in R^{N\times 1\times H\times W}$,级联拼接特征信息,经卷积核尺寸为 7 的卷积层将特征通道压缩为 1,再经 Sigmoid 函数激活后得到文本区域的空间注意力权重映射 $S_n(I_n)$。通道注意力在空间维度上压缩输入特征图 I_n,经全局平均池化和最大池化得到不同维度的空间背景特征:$F_g \in R^{N\times C\times 1\times 1}$ 和 $F_m \in R^{N\times C\times 1\times 1}$,通过共享的多层感知机(multilayer perceptron, MLP)网络进行非线性变换,输出两个不同的特征图逐点求和并利用 Sigmoid 函数激活,得到文本通道注意力映射 $C_n(I_n)$。$S_n(I_n)$ 和 $C_n(I_n)$:

$$S_n(I_n)=\text{sigmoid}(K^{7\times 7}([\text{AvgPool}(I_n)\oplus\text{Maxpool}(I_n)])) \tag{4}$$

$$C_n(I_n)=\text{sigmoid}(\text{MLP}(\text{AvgPool}(I_n))+\text{MLP}(\text{MaxPool}(I_n))) \tag{5}$$

式中,$K^{7\times 7}$ 表示 7×7 卷积,$\oplus$ 表示维度拼接操作,MLP 表示经过采用 ReLU 函数激活的多层感知机。

注意力机制优化了网络关注区域,而特征提取过程依赖于卷积层感受野。因脱机文本书写风格多变,使得受限于单一感受野的单卷积核,并不能较好适应文本变化。小尺寸卷积核注重于文本浅层特征,输入图片较大的情况下,需经过多层叠加以扩充感受野来整合高阶信息。较大尺寸卷积核全局信息提取能力更强,但忽视了细微特征,在深层特征语义信息下,影响

网络表达能力。基于上述特性,引入提供变化感受野的分支卷积。通过并行叠加不同尺寸的卷积核,利用不同的感受野来实现不同尺寸特征融合。同时结合多层感知机,达到高维空间非线性变换。多次叠加分支卷积构成注意力多分支卷积层 SeBlock,结构如图 7 所示。给定输入特征 $I_n \in R^{h\times w\times c}$,经注意力卷积层得到输出特征 $O_n \in R^{h'\times w'\times c'}$,$h'$、$w'$和 c'由卷积层决定,具体设置如表 1 所示。F_n 表示输入特征到输出特征的映射函数,即 $O_n = F_n(I_n)$,添加注意力的卷积块映射函数表示:

$$F_n(I_n) = (C_n[A_n(I_n)]+1)\times A_n(I_n)\times (D_C^{n_1\times n_1}+D_C^{n_2\times n_2}) \tag{6}$$

$$A_n(I_n) = [S_n(I_n)+1]\times I_n \tag{7}$$

式中,$D_C^{n\times n}$ 表示 $n\times n$ 深度分离卷积,$S_n(I_n)$ 表示空间注意力分支,$A_n(I_n)$ 表示空间注意力分支作用于输入特征,$C_n[A_n(I_n)]$表示通道注意力作用于施加空间注意力的特征。

记 $C_n' = C_n[S_n(I_n)\times I_n + I_n]$,$S_n' = S_n(I_n)$,将式(7)代入式(6),进一步展开得到:

$$F_n(I_n) = (C_n'\times S_n'\times I_n + C_n'\times I_n + S_n'\times I_n + I_n)\times D_C^{n\times n} \tag{8}$$

$$F_n(I_n) = [C_n'\times S_n' + C_n' + S_n')\times I_n + I_n]\times D_C^{n\times n} \tag{9}$$

式(9)由两个部分组成,第一部分表示注意力模块对上一层卷积主干提取的特征从不同方面进行映射,抽取细粒度特征。与卷积主干特征相乘,Sigmoid 函数会将特征值限制在 0 ~ 1,以此增强相关特征信息和抑制不相关特征信息。第二部分表示分支卷积作用过程,对注意力映射和残差连接的原始输入特征进行高层特征提取,以实现注意力引导卷积层。

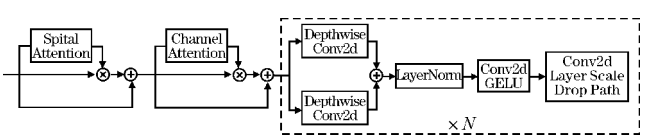

图 7 SeBlock 结构图

表 1 卷积层结构设置

层号	输出大小	操作
Convolution	H×W×1	K:7×7, S:2
SeBlock1	H/2×W/2×64	[7×7] + [3×3]
DownSample1	H/4×W/4×128	K:3×3,S:2
SeBlock2	H/4×W/4×128	[7×7]+[3×3]
DownSample2	H/8×W/8×256	K:3×3, S:2
SeBlock3	h/8×w/8×256	[5×5] + [3×3]
DownSample3	h/16×w/8×512	K:3×3, S:2×2
SeBlock4	h/16×w/8×512	[5×5] + [3×3]
DownSample4	h/32×w/8×1024	K:3,S:2×1
SeBlock5	h/32×w/8×1024	K:3, S:1

1.3.4 语义提取层

语义特征层首先使用时间卷积网络(temporal convolution network,TCN)做序列特征的提取,序列特征描述了文本的先后顺序,序列位置输出与序列之前位置有关,通过 TCN 提取感受野范围内的局部上下文信息,以学习序列依赖信息,引入空洞卷积扩大感受野,整合长距离的信息。

为了让序列信息有效融合,选择多头注意力机制[11]对序列特征进行自注意力计算,得到含有权重的特征表示,以此交互上下文信息。Transformer 编码器使用自注意力机制,输入序列中任意两个位置之间的距离缩小为一个常量,以键值对的形式建模输入序列间的语义关系,多头机制映射至不同的子空间去学习特征,优化不同特征部分。编码器结构如图 8 所示。

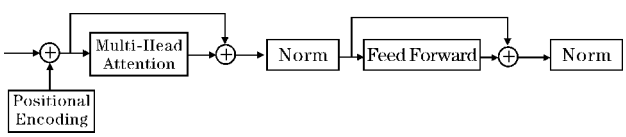

图 8 Transformer 编码器结构图

序列特征多头自注意力计算如下:

特征序列 $f_{\text{in}} \in R^{(N,L,C)}$ 经过权重矩阵映射得到 Q,K,$V \in R^{(N,L,C)}$,被 M 个注意力头均分为 Q_i,K_i,$V_i \in R^{(N,M,L,C/M)}$,映射过程如下:

$$Q = W^Q f_{\text{in}},\ K = W^K f_{\text{in}},\ V = W^V f_{\text{in}} \tag{10}$$

每个注意力头内通过 Q_i 与 K_i 的转置做点积运算,经过 Softmax 归一化,求得各个位置在序列中的不同关联程度 $\text{Att}_i \in R^{(N,M,L,L)}$:

$$\text{Att}_i(Q_i,K_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{11}$$

式中,$\sqrt{d_k}$ 为缩放因子,缩放内积避免过大,d_k 为 K 的维度。

得到的权重向量再与 V 做点积,加权各位置语义输出,以此融合不同位置的语义特征,再拼接上不同头输出,形成多头注意力,计算如下:

$$\text{head}_i = \text{Att}_i \cdot V$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\cdots,\text{head}_n)W^o \tag{12}$$

2 实验结果与分析

2.1 数据集与参数设置

实验所用的数据集为公开数据集 CASIA-HWDB2.x,由 1019 名书写者书写完成,包含 5091 张文本图片,2703 类字符类别,字符数为 1349414。数据集划分为训练集和测试集,其中 4076 张图片用作训练

集,1015 张用于测试集,

实验平台为 Win10 操作系统、NVIDIA GeForce RTX 3070 显卡,使用 PyTorch 框架搭建网络,优化器选用 Adam,初始学习率设为 0.001,批大小数量设置为 8,训练轮数 epoch 为 50。文本图像大小调整为 736×736 输入网络中,不额外使用语言模型来优化识别结果。

为评估识别网络的性能,采用 Levenstein 字符编辑距离[12]用作度量方式,计算插入、删除和替换的字符数,得到准确率(AR)和正确率(CR)两个评价指标,计算如下:

$$\begin{aligned} \mathrm{AR} &= (N_c - D_c - I_c - S_c)/N_c \\ \mathrm{CR} &= (N_c - D_c - S_c)/N_c \end{aligned} \tag{13}$$

其中,N_c 表示输入文本图像的字符序列长度,D_c 表示需要删除字符的数量,I_c 表示需要插入字符的数量,S_c 表示替换错误字符的数量。

2.2　实验对比及分析

近年来不同方法在 CASIA-HWDB2.x 数据集上识别性能的对比如表 2 所示,分别提供了不同条件下识别精度结果。文献[5]使用多维长短时记忆循环网络 MDLSTM-RNN 结合 CTC 链接时序分类用于端到端识别,文献[14]和文献[15]使用 CNN 提取特征,并在 LSTM 和 CNN 分别引入注意力机制赋予特征不同的权重,识别性能较文献[5]取得很大提升,意味着注意力机制能增强网络对重要信息的捕捉能力。文献[13]使用 CNN-ResLSTM 结合数据预处理以及文本图片纠正,文献[16]使用像素级别纠正的深度网络进行 CNN 和 RNN 中像素纠正,识别率分别提升1.78%和4.4%,说明文本纠正有助于识别提升。此外,文献[5]、[13]和文献[15]额外使用语言模型以提高识别准确率。

表 2　不同方法识别结果　　单位:%

方法	无语言模型		有语言模型	
	AR	CR	AR	CR
MDLSTM+CTC[5]	–	93.59	–	96.63
CNN + ResLSTM + CTC[13]	94.90	95.37	96.97	97.28
CNN + LSTM + Attention[14]	95.76	96.73	–	–
ResGate[15]	96.85	97.46	97.32	97.90
Recogm + Rectm[16]	97.31	97.90	–	–
本文方法(全文)	96.84	97.52	–	–
本文方法(文本行)	97.53	98.00	–	–

本文所提方法在特征提取阶段使用注意力多分支卷积,提供变化感受野,融合不同尺度特征,语义提取阶段使用自注意机制构建序列特征语义上下文,因而具有更好的特征提取能力。除文献[16]和使用语言模型的文献[15]外,表 2 中其余方法 CR 准确率均低于本文方法。由于参与对比的方法皆为单文本行输入图片识别结果,针对本文方法有效性讨论,额外测试单文本行输入图片下的识别性能,结果如表 2 最后一项,所提方法取得 CR 和 AR 较最高 CR 和 AR 准确率皆有提升,验证了本文模型的可行性。

另外,本文还在 CASIA-HWDB2.x 数据集上进行一系列消融实验,以验证所提模型的有效性。首先对注意力卷积层特征提取能力分析,再在最终识别模型的基础上删除注意力卷积层的不同组件,性能对比如表 3 所示。注意力和多分支特征融合在单独使用下,CR 分别提升0.5%和0.2%,这意味着注意力分支卷积层提取特征能力更强,赋予网络更强的泛化性。

表 3　注意力卷积层组件结果对比　　单位:%

模块	AR	CR
无	95.9	96.7
+Attention	96.5	97.2
+ multi kernel	96.1	96.9

模型选用 TCN 和 transformer 编码器作为序列和语义特征提取层,该模块由两部分所构成。为了验证不同部分对性能的影响,对不同配置获得的精度和速度进行了比较,所有的实验都在同一个数据集和特征提取网络下进行,实验结果如表 4 所示。

表 4　TCN 和 Transformer 堆叠层数对比结果

模块	AR/%	CR/%	时间/(ms/张)
无	94.7	95.3	196
+ TCN	94.9	96.2	197
+ Transformer×4	95.9	96.4	199
+ Transformer×6	96.1	96.6	200
+ Transformer×8	96.1	96.5	202
+TCN+Transformer×4	96.7	97.3	204
+TCN+Transformer×8	96.8	97.4	211

可以看出,使用 TCN 和 Transformer 提升了网络精度,CR 和 AR 在 TCN 和 6 层 Transformer 的配置达到最高,而随着 Transformer 层数的不断加深,准确率有所下降,可能深度过深引起网络退化;单张图片推理时间由204 ms增长至211 ms,较不使用语义提取层,推理时间增加7.5%,TCN 的使用对推理时间几乎无影响,表明语义提取层对推理速度无明显降低。

图 9 为模型的训练曲线,图 9(a)和(b)为训练损失以及验证损失值曲线。随着训练轮数 epoch 增加,损失值快速下降,20 轮后曲线趋近于平稳。图 9(c)和(d)为 CR 和 AR 准确率曲线,两者总体趋势趋近于一致,快速上升后缓慢增长。

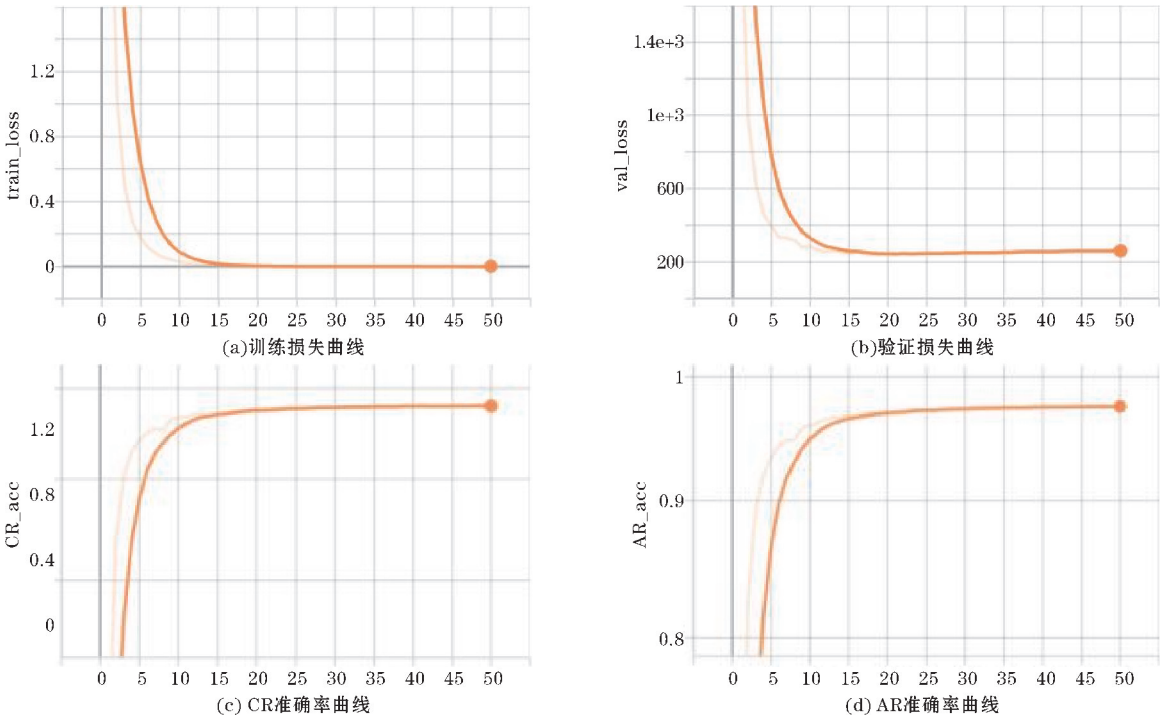

图 9 网络训练曲线图

3 结束语

针对手写中文文本识别,提出一种注意力分支卷积和 Transformer 的文本定位和识别方法。文本识别网络利用透视变换将文本图像定位信息进行倾斜文本纠正;特征提取阶段使用注意力分支卷积获取文本区域的注意力分布和变感受野特征融合,从而有效适应长文本的变化;语义提取层使用 TCN 和 Transformer 用于整合序列特征和上下文语义特征提取。在公开数据集上进行实验,结果表明所提方法的可行性。接下来的研究工作将应用于其他手写体语言。

参考文献:

[1] Kumar M, Jindal M, Sharma R. Segmentation of isolated and touching characters in offline handwritten Gurmukhi script recognition[J]. International Journal of Information Technology Computer Science, 2014, 6(2): 58-63.

[2] Qiufeng Wang, Fei Yin, Chenglin Liu. Handwritten chinese text recognition by integrating multiple contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1469-1481.

[3] Su T H, Zhang T W, Guan D J, et al. Offline recognition of realistic Chinese handwriting using segmentation-free strategy[J]. Pattern Recognition, 2009, 42(1): 167-182.

[4] 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 Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(11): 2298-2304.

[5] Ronaldo Messina, Jerome Louradour. Segmentation-free handwritten Chinese text recognition with LSTM-RNN[C]. International Conference on Document Analysis and Recognition, 2015: 171-175.

[6] Yichao Wu, Xiaolin Hu. From Textline to Paragraph: A promising practice for Chinese text recognition[C]. Proceedings of the Future Technologies Conference, 2020: 618-633.

[7] 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.

[8] Wang W. PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 9: 5349-5367.

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference

on Computer Vision and Pattern Recognition (CVPR),2016:770-778.

[10] 张宸嘉,朱磊,俞璐. 卷积神经网络中的注意力机制综述[J]. 计算机工程与应用,2021,57(20).

[11] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: MIT Press,2017:5998-6008.

[12] Vladimir I Levenshtein. Binary codes capable of correcting deletions,insertions,and reversals[J]. Soviet Physics Doklady,1966,10(8):707-710.

[13] Xie C,Lai S,Liao Q,et al. High Performance Offline Handwritten Chinese Text Recognition with a New Data Preprocessing and Augmentation Pipeline[C]. Document Analysis Systems. DAS 2020. Lecture Notes in Computer Science,2020:12116.

[14] 王馨悦,董兰芳. Attention 机制在脱机中文手写体文本行识别中的应用[J]. 小型微型计算机系统,2019,40(9):1876-1880.

[15] Yintong Wang, Yingjie Yang, Weiping Ding, et al. A residual-attention offline handwritten Chinese text recogni-tion based on fully convolutional neural networks [J]. IEEE Access, 2021, 9: 132301-132310.

[16] Xiao S,Peng L,Yan R,et al. Deep Network with Pixel-Level Rectification and Robust Training for Handwriting Recognition[C]. International Conferen-ce on Document Analysis and Recognition (ICDAR),2019:9-16.

Handwritten Text Recognition based on Attentional Multi-branch Convolution and Transformer

ZHENG Xiaoxu, SHU Shanshan, WEN Chengyu

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

Abstract:The handwriting recognition technology has been widely studied as a key part of the automatic paper marking. A handwritten Chinese text recognition method for text localization and recognition is proposed for the problem of complex handwriting of Chinese handwritten text. The text localization information is corrected by using perspective change for skewed text, followed by feature extraction stage using attentional multi-branch convolutional layer to extract key region features of text images and multi-scale feature fusion, semantic extraction stage by constructing sequence information and modeling contextual semantic information through temporal convolutional network and Transformer encoder, and finally by connecting temporal classification functions to achieve sequence features and character sequence label alignment. The proposed method is investigated using the publicly available dataset CASIA-HWDB, and the results show that the attention branching convolutional layer and the semantic extraction layer can effectively improve the algorithm performance, which verifies the feasibility of the proposed method.

Keywords:handwriting text recognition;Transformer;attention mechanism;connectionist temporal classification