

基于深度学习的甲骨文偏旁与合体字的识别研究

林小渝¹, 陈善雄¹, 高未泽¹, 莫伯峰², 焦清局³

(1. 西南大学计算机与信息科学学院, 重庆 400715)

(2. 首都师范大学甲骨文研究中心, 北京 100048)

(3. 安阳师范学院 计算机与信息工程学院, 河南 安阳 455000)

[摘要] 由于甲骨文字形结构多样, 异体字较多, 其识别一直是甲骨文领域研究的重要问题. 本文首次提出以甲骨文偏旁为识别的基本构件, 建立单偏旁和合体结构的甲骨文字符识别方法, 提升甲骨文识别的精度. 方法一: 根据甲骨文偏旁字形特点, 对甲骨文拓片上的合体字进行甲骨文单偏旁最大极值稳定区域的选取, 然后, 通过改进的 BN-LeNet 模型识别甲骨文各个偏旁; 方法二: 针对甲骨文合体字拓片稀缺的问题, 本文提出一种直接对甲骨文合体字进行整体识别的 OraNet 模型, 该模型采用迁移学习的训练策略, 对在脱机手写汉字 HCL2000 数据集预训练的卷积神经网络模型进行参数和结构上的微调, 实现迁移得到低层表示和甲骨文合体字集上高层表示的特征融合, 以此来提取甲骨文合体字的高级特征. 实验结果表明, BN-LeNet 网络对甲骨文单偏旁识别率为 96.24%, 微调的 OraNet 模型对甲骨文合体字识别率为 98.58%, 从而表明从甲骨文单偏旁的角度进行甲骨文字形识别, 可以获得较高的识别精度. 同时本文将甲骨文视为偏旁组合而非整字识别, 这使得算法能够识别从未见过的甲骨文新字, 即零样本学习, 对甲骨文研究有着重要的应用意义.

[关键词] 甲骨文识别, 甲骨文偏旁, 迁移学习, 偏旁分析, 非极大值抑制算法

[中图分类号] TP399 **[文献标志码]** A **[文章编号]** 1001-4616(2021)02-0104-13

Oracle Radical and Oracle Combined Character Recognition Based on Deep Learning

Lin Xiaoyu¹, Chen Shanxiong¹, Gao Weize¹, Mo Bofeng², Jiao Qingju³

(1. School of Computer and Information Science, Southwest University, Chongqing 400715, China)

(2. Oracle Research Center, Capital Normal University, Beijing 100048, China)

(3. School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China)

Abstract: Due to the complex glyph structure and many variants, the recognition of oracle characters have always been an important problem in relevant field. This paper proposes to use oracle radical as a component to establish a recognition method of oracle radical and oracle combined character, to improve the accuracy of oracle bone script recognition. Method 1: According to the characteristics of the oracle radical, we have selected the maximum extreme stable region (MSER) of the single radical from the oracle bone script, and then put it into the improved BN-LeNet model to recognize; Method 2: In view of the scarcity of oracle combined character, we have proposed an OraNet model that directly recognizes the oracle bone script character. Transfer learning was introduced into model training, so as to extract the high-level features of the oracle bone script, the fine-tuning strategy is implemented to achieve the feature aggregation of low-level representations and high-level representations. The experimental results show that the recognition rate of BN-LeNet network for oracle radical recognition is 96.24%, and the recognition rate of fine-tuned OraNet model for oracle combined character is 98.58%, which shows that considering the oracle bone script recognition from the perspective of oracle radical, higher recognition accuracy can be obtained. At the same time, we treat oracle bone script as a radical combination instead of whole-word, which enables the system to recognize unseen new oracle bone script, i.e., zero-shot learning. so it has important application significance for Oracle research.

Key words: oracle character recognition, oracle radical, transfer learning, radical analysis, non-maximum suppression

收稿日期: 2020-10-18.

基金项目: 国家社会科学基金项目 (19BYY171)、甲骨文信息处理教育部重点实验室项目 (OIP2019E009).

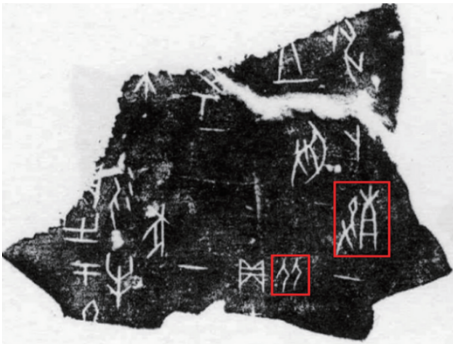
通讯作者: 陈善雄, 博士, 副教授, 研究方向: 模式识别与图像处理. E-mail: 304904212@qq.com

从 1899 年甲骨文发现至今,已有 120 多年的历史,甲骨文以其自身独特的历史文化价值引起了众多学者的研究兴趣^[1],现已发展成为一门重要学科——甲骨学. 2019 年 11 月 2 日,在纪念甲骨文发现和研究 120 周年座谈会上宣读了习近平总书记的贺信^[2],贺信指出:“甲骨文是迄今为止中国发现的年代最早的成熟文字系统,是汉字的源头和中华优秀传统文化的根脉,值得倍加珍视,需多部门多学科协同开展甲骨文的研究和应用.”对此,我们要综合运用人工智能识别等技术手段,让甲骨文实现更好传承与发展.

100 多年以来,先后有 15 万余片甲骨文出土,历经几代学者的共同努力,认识或基本认识的甲骨文字却仅占总数的 2/5,甲骨文考释仍然困难重重. 总结起来主要有两方面原因:一是时间因素. 已知最早的商代甲骨文,距今已经有 3 600 多年,在这么长的时间里,汉字的形体、意义和读音都发生了很大的变化,这些变化都是甲骨文考释的障碍. 二是空间因素. 甲骨文是众多刻写者(刻手)留下的文字材料,而当时缺乏十分明确的字形规范,即使同一甲骨文字,也存在着多种不同的形体(异体字),这种变化更增添了甲骨文考释的难度.

尽管考古学家和古文字学家不断努力,但从计算机视觉角度对甲骨文文字分析的研究却较少,也缺乏相关的公共数据集. 虽然已有一些传统的方法尝试对甲骨文字进行识别,并且取得了一定的成果,如顾绍通基于拓扑配准识别甲骨文的方法^[3];周新伦等基于图论和笔划特点识别的方法^[4];李锋等基于图特征来识别甲骨文的方法^[5];栗青生等基于图同构的方法来识别甲骨文字形的方法^[6]. 这些研究的共同特征是将甲骨文字符当作一个整体图形来处理,而没有考虑不同字符之间的相似性和内部结构. 由于甲骨文中合体字情况大量存在(如图 1 所示),整体识别一些固有的局限性使得上述研究都存在发展的壁垒,因而实用性受到限制.

为此,本文从甲骨文单偏旁的角度出发,分析组成甲骨文常用的基本结构单位——偏旁,以此来进行甲骨文的识别. 一方面,本文建立了甲骨文常见的单偏旁数据集,在此基础上设计甲骨文单偏旁自动识别的算法,方便考古学家和古文字学家在考释甲骨文新字时,利用单偏旁分析辅助考释. 另一方面,对于甲骨文字符识别领域来说,一个巨大的难题就是特定领域数据的缺乏,特别是在现有甲骨文字库中的甲骨文合体字数. 因此,本文在建立的甲骨文单偏旁数据集的基础上,利用计算机辅助拼接技术生成了数量庞大的甲骨文合体字数据集,如图 2、3 所示.



红色标记为甲骨文合体字

图 1 龟甲拓片上的甲骨文字符

Fig. 1 A turtle shell with multiple oracle instances carved on it



图 2 左右结构不同变体的甲骨文拼合图

Fig. 2 Different variants of the same character (left and right structure)

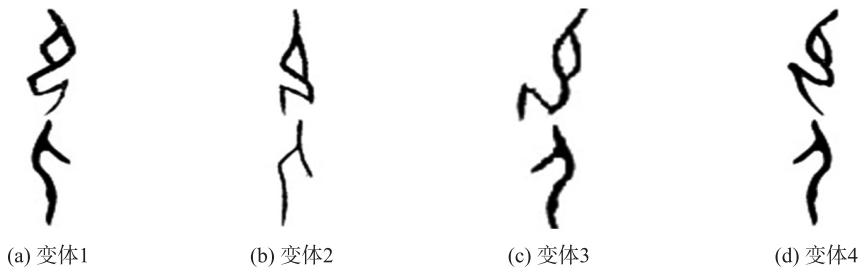


图 3 上下结构不同变体的甲骨文拼合图

Fig. 3 Different variants of the same character (upper and lower structure)

此外,由于新见的甲骨文字符必然超出现有数据集的范围,因此识别这些甲骨文字符通常是一个零样本学习问题. 近几年,零样本学习(zero-shot learning)引起了研究人员的浓厚兴趣^[7],这种类型的学习是一个具有挑战性的问题,但具有巨大的潜在价值,所以本文提出从甲骨文偏旁分析来考释甲骨文字符.

本文第一个贡献在于甲骨文单偏旁检测识别框架的设计. 由于我们缺乏甲骨文单偏旁标注数据集,深度学习检测算法并不适用于此. 受汉字偏旁拆分的启发,本文采用最大极值稳定区域(MSER)算法来筛选甲骨文单偏旁文本区域,再采用非极大值抑制算法进行后处理. 在甲骨文单偏旁识别实验过程中,由于甲骨文单偏旁字符形状不规则,单偏旁变形较多,导致甲骨文单偏旁特征难以表征,传统汉字文字识别方法未能解决该问题,因此本文设计了 BN-LeNet 网络,对有 d 维的 $x=(x^{(1)}\cdots x^{(d)})$ 的输入数据,对每个批次的每一个维度进行归一化处理,以此解决数据分布改变和网络收敛较慢的问题,并在网络结构中插入 Dropout 层和 L2 正则化防止模型过拟合,可以有效地学习甲骨文拓片中有判别意义的深度特征.

本文第二个贡献在于甲骨文合体字识别框架的设计. 由于甲骨文合体字图像稀缺,我们进行了大量的实验来评估不同因素对甲骨文合体字识别精度的影响,并与传统 CNN 网络在甲骨文字合体字识别比较,实验结果证明本文甲骨文合体字识别方法的有效性. 近期研究表明^[8],当外部数据可用时,对辅助任务进行预训练,然后对特定领域的数据进行微调,是学习复杂领域特定模型的有效范例. 基于此,本文利用卷积神经网络在脱机手写汉字 HCL2000 数据集^[9]上预训练,在结合本文建立的甲骨文合体字数据集进行联合训练,微调顶部三层卷积层提取深度特征,使卷积特征更有利于甲骨文合体字的表示,以解决在甲骨文合体字样本稀缺的情况下,模型泛化能力差等问题,使模型对预测甲骨文合体字图像具有更好的鲁棒性.

1 模型设计

1.1 甲骨文单偏旁的检测和识别

与英语或阿拉伯文字不同,汉字从甲骨文开始就可以分解成有限数量的偏旁,这些偏旁及其相互间的位置关系都具有区别字形的功能. 现有的文本识别模型大多数是面向现代汉字的,而甲骨文等古文字并不具有现代汉字丰富的数据集,所以针对现代汉字非常有效的整体识别方案,并不能简单地套用在甲骨文识别上,基于这一认识,本文从甲骨文偏旁角度进行甲骨文识别的研究. 基于偏旁的甲骨文字符识别主要包括偏旁提取和偏旁识别两个方面,本文先将甲骨文字符拆分成多个单偏旁,再对各个单偏旁顺序识别,最后通过结构标记组成该字符识别结果,图 4 是本文提出的甲骨文单偏旁检测识别框架流程.

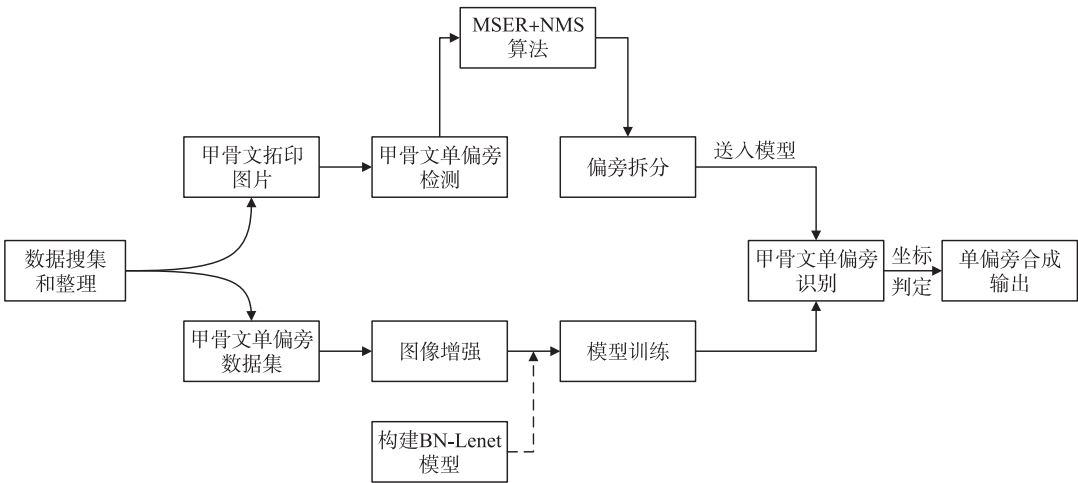


图 4 甲骨文单偏旁检测识别框架流程图

Fig. 4 The Oracle's single radical detection and recognition framework by our proposed approach

1.1.1 甲骨文单偏旁区域检测

在甲骨文拓片中,由于背景和文本区域灰度对比度较大,而甲骨文单偏旁文本区域内部灰度几乎不变,并且该区域能够在多重阈值下保持形状不变,符合最大极值稳定区域的特性,因此利用这一特性,本文采用最大极值稳定区域(MSER)算法来筛选甲骨文单偏旁文本区域.

甲骨文单偏旁 MSER 选取具体实现过程如下:

将甲骨文图像转换成灰度图像,在灰度区间 $[0, 255]$ 内的 256 个不同阈值对灰度图像进行二值化,令 Q_i 表示二值化阈值 i 对应的二值图像中的某一连通区域,当二值化阈值由 i 变成 $i+\Delta$ 和 $i-\Delta$ 时, Δ 为变化值,连通域 Q_i 相应变成了 $Q_{i+\Delta}$ 和 $Q_{i-\Delta}$.

随着亮度阈值的增加或者减少,计算公式(1),当两个不同阈值间的区域变化不超过阈值 i 就被认为是稳定的.

$$v(i) = \frac{|Q_{i+\Delta} - Q_{i-\Delta}|}{|Q_i|}, \quad (1)$$

式中, Q_i 表示第 i 个甲骨文单偏旁连通区域的面积, Δ 表示微小的阈值变化,当 $v(i)$ 小于给定阈值时认为该甲骨文单偏旁区域为最大极值稳定区域.

为了进一步对 MSER 得到的甲骨文单偏旁不规则区域进行处理,需要对选定的甲骨文单偏旁区域进行轮廓拟合.

甲骨文单偏旁 MSER 区域轮廓拟合实现过程如下:

对甲骨文单偏旁最大极值稳定区域内的每个点,计算整个甲骨文单偏旁极值稳定区域的几何 0 阶矩和几何 1 阶矩:

$$m_{00} = \sum I(x, y) \quad m_{01} = \sum yI(x, y) \quad m_{10} = \sum xI(x, y), \quad (2)$$

得到整个甲骨文单偏旁极值稳定区域的重心位置 $x_c = \frac{m_{10}}{m_{00}}, y_c = \frac{m_{01}}{m_{00}}$.

确定甲骨文单偏旁区域轮廓拟合的长半轴、短半轴、角度(长半轴与 x 轴顺时针):计算中心二阶矩 $U =$

$$\begin{bmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{bmatrix}.$$

式中

$$\begin{aligned} \mu_{20} &= \sum (x - x_c)^2 I(x, y), \quad \mu_{02} = \sum (y - y_c)^2 I(x, y), \\ \mu_{11} &= \sum (x - x_c)(y - y_c) I(x, y). \end{aligned} \quad (3)$$

计算该二阶矩的两个特征值,有

$$\begin{aligned} \lambda_1 &= \frac{(\mu_{20} + \mu_{02}) + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2}, \\ \lambda_2 &= \frac{(\mu_{20} + \mu_{02}) - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2}, \end{aligned} \quad (4)$$

于是可以分别得到甲骨文单偏旁区域轮廓拟合的长半轴、短半轴、角度

$$\begin{aligned} a &= \sqrt{\frac{\lambda_1}{m_{00}}} \quad b = \sqrt{\frac{\lambda_2}{m_{00}}}, \\ \theta &= \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right). \end{aligned} \quad (5)$$

在选取了甲骨文单偏旁最大极值稳定区域后,为了进一步提升检测精度,本文在 MSER 基础上加入 NMS 后处理,筛选最优甲骨文单偏旁区域,在实验过程中发现,在对甲骨文单偏旁进行 MSER 检测之前,先进行形态学腐蚀操作,可以提升检测精度,图 5 是检测结果示例图.

1.1.2 甲骨文单偏旁识别

由于甲骨文单偏旁字符形状不规则,单偏旁变形较多,导致甲骨文单偏旁特征难以表征,本文构造了包括批归一化层、Dropout 层、正则化的模型用于甲骨文单偏旁的识别,称为 BN-LeNet 网络模型,如图 6 所示. 模型由 2 个卷积层、2 个最大池化层、1 个全连接层、1 个 softmax 层、3 个批归一化层和 1 个随机失活层构成.

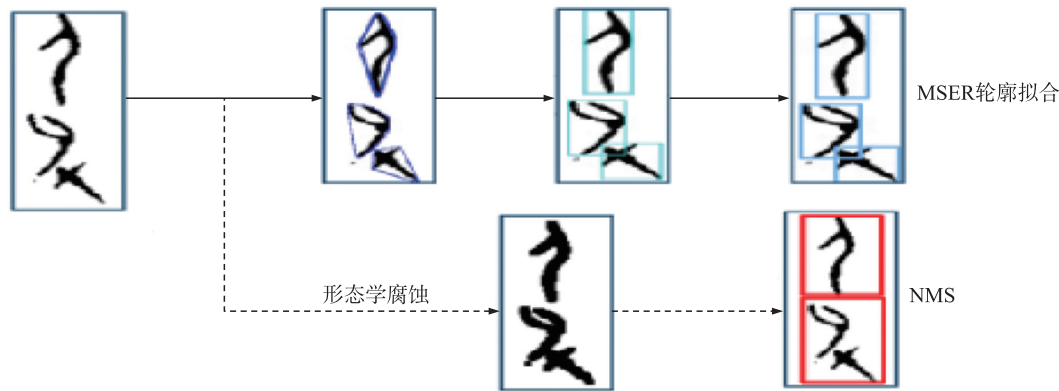


图 5 MSER+NMS+形态学腐蚀结果图
Fig. 5 MSER+NMS+Morphological erosion results

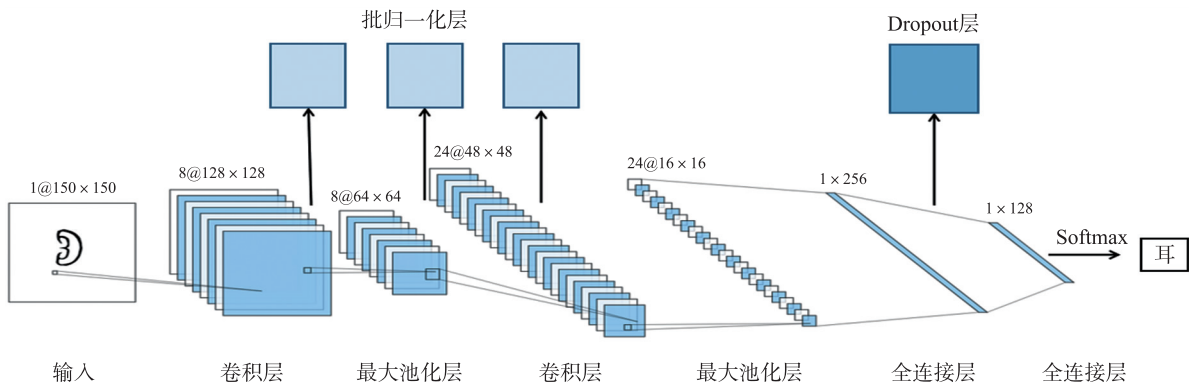


图 6 BN-LeNet 网络模型图
Fig. 6 BN-LeNet network model

为了解决数据分布改变和模型过拟合的问题,我们在网络的卷积层和池化层中间插入一个批归一化层,总共添加 3 个批归一化层,即对有 d 维的 $x = (x^{(1)} \cdots x^{(d)})$ 的输入数据,对每个批次的每一个维度进行归一化处理.

$$\hat{x}^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{V[x^{(k)}]}}, \quad (6)$$

上式中, $E[\cdot]$ 和 $V[\cdot]$ 是通过甲骨文单偏旁训练集计算得到. 如果对某一层输出数据只做归一化处理,然后送入网络的下一层,这样会影响网络该层所学习到甲骨文单偏旁特征. 因此对网络层的每一个激活函数 $x^{(k)}$ 引入一组可学习参数 $\gamma^{(k)}$ 和 $\beta^{(k)}$ 进行变换重构

$$y^{(k)} = \gamma^{(k)} \hat{x}^{(k)} + \beta^{(k)}. \quad (7)$$

$\gamma^{(k)}$ 和 $\beta^{(k)}$ 和原始模型的参数一起学习,且通过设置 $\gamma^{(k)} = \sqrt{V[x^{(k)}]}$, $\beta^{(k)} = \sqrt{E[x^{(k)}]}$ 时可以恢复某一层学到的原始甲骨文单偏旁特征.

为了防止模型过拟合,通过实验对比,我们在 Conv1 卷积层上添加 $L2$ 正则,以及在全连接层和 Softmax 层添加概率 $p=0.5$ 的 Dropout 层^[10],在每个训练 epoch 中,在前向传播的时候,让某个神经元的激活值以一定的概率 p 停止工作,可以明显地减少模型过拟合现象.

1.2 甲骨文合体字的整体识别

为了解决甲骨文合体字样本量小的问题,一方面我们将各单偏旁拼合成甲骨文合体字作为训练集,另一方面考虑脱机手写汉字 HCL2000 数据集与甲骨文合体字相似性高,预训练网络卷积层学到的特征空间层次结构适用于甲骨文合体字识别,例如底层提取的局部且高度通用的图像边缘、笔画纹理等特征,考虑到 Vgg-16 模型在特征提取和模型结构的优势,本文选择 Vgg-16 网络作为特征提取的基网络. 由于脱机手写汉字 HCL2000 数据集中为手写汉字图像集,虽部分笔画特征和甲骨文合体字特征相似,但直接用手写汉字的特征进行甲骨文识别,效果不佳,于是将在脱机手写汉字 HCL2000 数据集上预训练好的 Vgg-16

模型进行微调,先冻结预训练好的 Vgg-16 模型,然后解冻顶部三层卷积层,再在建立的甲骨文合体字数据集上进行联合训练,以便让卷积基提取的特征表示与甲骨文合体字特征更加相关,再把原网络最后一个包含 1 000 个神经元的全连接层替换为包含 450 个神经元的全连接层,以此适应甲骨文合体字的分类任务,图 7 是 OraNet 框架图。

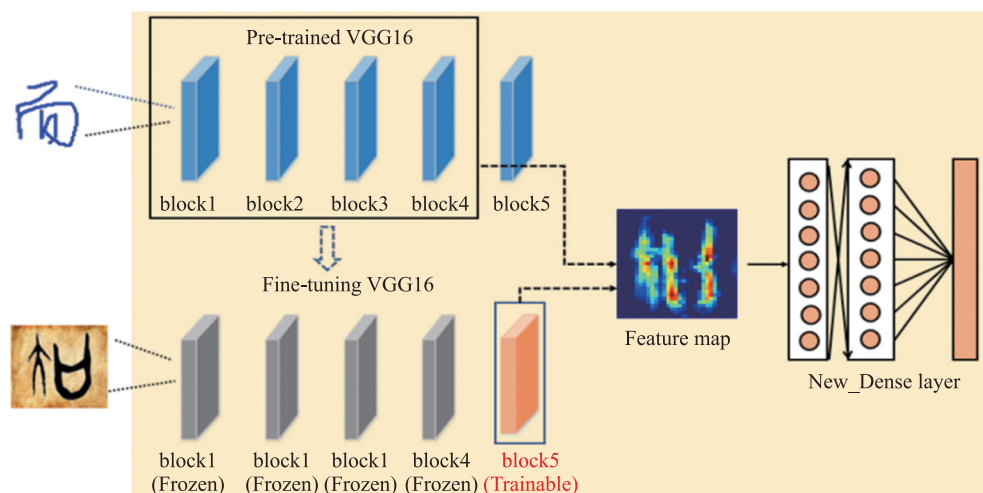


图 7 OraNet 框架图

Fig. 7 The framework of OraNet

甲骨文合体字识别的训练策略:

- (1) 将 Vgg-16 模型脱机手写汉字 HCL2000 数据集上进行预训练得到 Conv_base 网络,去除网络的密集连接分类器层。
- (2) 在顶部添加 Dense 层来扩展 Conv_base 网络,分类器类别数设为甲骨文合体字类别数 450。
- (3) 冻结卷积基的前 4 个卷积块,以此来防止卷积基在脱机手写汉字 HCL2000 数据集上学到的表示不被修改。
- (4) 解冻 Conv_base 的最后一个卷积块,并在在甲骨文合体字数据集上训练 block5。
- (5) 将前 4 个卷积块的输出特征 F_{block4} 和卷积块 5 输出特征 F_{block5} 进行融合成高层特征 F_{oracle} 输入 Dense 层进行分类。

$$F_{oracle} = [\alpha \times F_{block4}, \beta \times F_{block5}], \quad (8)$$

式中, α 、 β 为权重参数,本文分别取值为 0.6、0.4。为了较为直观地展示模型提取特征,本文借助 Jupyter Notebook 可视化卷积层特征图,如图 8 所示,网络各个卷积块对于甲骨文合体字不同位置的响应值不同。

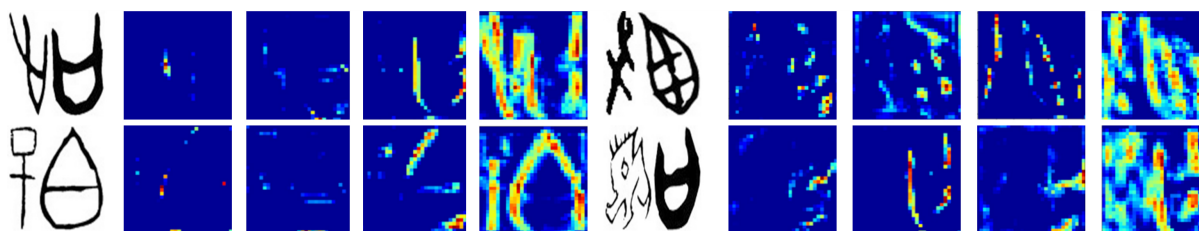


图 8 网络各卷积块输出特征图

Fig. 8 Output feature map of each convolution block of the network

2 数据集

2.1 数据采样

本文所使用的甲骨文单偏旁样本由甲骨文研究中心的专家提供,这些样本来自于人工拓印所得,该样本库包括多个甲骨文单偏旁,每个偏旁约有 20 多个变形,同时为了克服手写样本的不足,本文设计了甲骨文样本采集表,每页 123 个甲骨文单偏旁,每行 14 个甲骨文单偏旁,共 9 行,奇数行为人工拓印图,偶数行

为人工临摹写得,共组织 22 人临摹,每人 2 页,共得到 5 412 个甲骨文手写单偏旁,样表如图 9 所示。

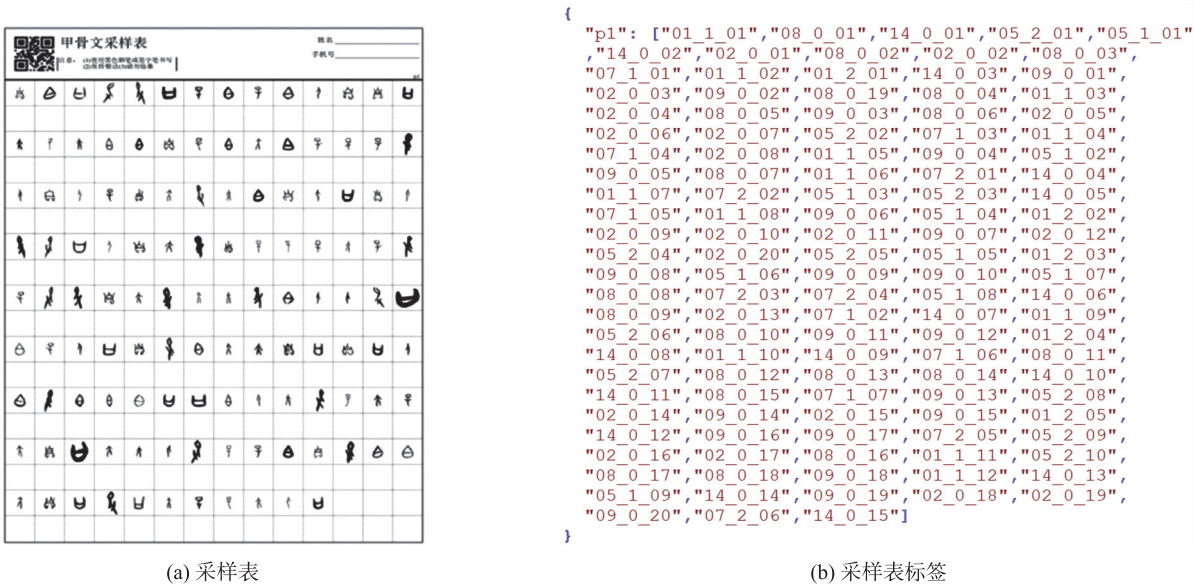


图 9 采样表及数据标签
Fig.9 Oracle radical sampling table and label

本文将样本表进行图像切割,并且还对甲骨文单偏旁数据集进行了数据扩增,对图像进行旋转 30°/330°、裁剪、上下左右 4 个方向各平移 30%等操作来扩增数据集,同时在甲骨文单偏旁数据集的基础上,利用拼接技术建立了甲骨文合体字样本库,部分示例如图 2、3 所示。

3 实验结果与分析

3.1 实验设置

实验环境:Inter i7-7700 CPU,显卡 NVIDIA GeForce GTX 2080,内存 16GB,通过 PyCharm 编程环境和深度学习 Keras 框架进行实现,并借助 TensorBoard 和 matplotlib 等可视化工具对训练输出进行可视化,数据集中的甲骨图片源于甲骨文研究相关论著的配图和对甲骨片的扫描图像,由于原始的图像大小不一,成像质量参差不齐,且有噪声干扰,本文对所有图像进行了处理,消除了噪声,为了避免数据不一致带来实验结果的误差,本文工作中相关的实验均使用相同的数据格式,即调整所有图像大小为 150×150 像素,并将增强后的甲骨文单偏旁数据集和甲骨文合体字数据集随机分成两部分:60%作为训练集用于模型拟合的数据样本;40%作为测试集用于评估模最终模型的泛化能力。

3.2 评价指标

本文中使用了 MSER+NMS 方法对甲骨文单偏旁进行检测,我们采用传统文本检测相关评价指标,准确率、召回率, $hmean(f-score)$ 作为综合性能评价指标。

对于甲骨文单偏旁图像分类和甲骨文合体字图像分类,我们从图像的分类准确率来判断该分类网络的性能,令代表测试集中图像的总数量,代表其中分类正确的图像数量,则分类准确率可以表示为:

$$Accuracy = \frac{N_{rec}}{N_{total}} \quad (8)$$

3.3 甲骨文单偏旁检测对比实验分析

对于甲骨文单偏旁检测,由于图像标注成本代价较高,这里没有对甲骨文单偏旁数据集进行图像标注,所以采用传统文本检测算法,为了对比不同的传统检测方法对甲骨文单偏旁检测精度的影响,设置了 3 组实验。

实验 1:为了比较传统检测方法对甲骨文单偏旁的检测效果,对传统的文本检测算法如笔画宽度变换(SWT)文本检测算法、传统投影算法、连通分量算法、MSER 文本检测算法等进行了比较。

实验 2:为了验证 NMS 的必要性,进行了两个方面的测试:(1)仅使用 MSER 算法;(2)采用 MSER+

NMS 算法检测.

实验 3:为了验证形态学腐蚀操作的必要性,这里验证了仅用 MSER+NMS 方法能否达到较好的检测精度,进行了如下实验:(1)仅用 MSER+NMS 算法检测;(2)利用 MSER+NMS+形态学腐蚀算法检测.

从表 1 结果可以看出,通过对比七种不同检测方法的检测结果数据,可以发现,基于传统投影法的检测时间较短,但是准确率和召回率较低,主要是由于有些甲骨文上下结构和左右结构有字符区域重叠的现象,而基于连通域的检测方法能够较好的处理字符区域重叠的问题,但是对于粘连字符,还是不能取得较好的检测效果,在对两种传统的方法做了结合以后,准确率和召回率得到了一定的提升,基于 MSER 和 SWT 结合的方法使得效果下降,主要是因为甲骨文手写拓片的笔画宽度不一致,因此准确率反而有所下降,而且所用的运行时间也较长,基于 MSER+NMS 结合的算法,可以在 MSER 上过滤掉多余的框,从而使得准确率上升,加上本文提出基于 MSER+NMS+形态学腐蚀的方法,可以较好地处理字符区域重叠和字符粘连的情况,准确率和召回率取得最好的效果.

表 1 本文方法和其他传统检测方法在总体性能上的对比

Table 1 Comparison of the overall performance on this method with other traditional detection methods					
组数	传统文本检测算法	<i>P</i>	<i>R</i>	<i>H</i>	运行时间
实验 1	基于连通分量法 ^[11]	0.83	0.81	0.82	0.74
	基于传统投影法 ^[12]	0.78	0.71	0.74	0.37
	基于连通分量法+传统投影法 ^[13]	0.84	0.76	0.80	0.89
	基于 SWT 的方法 ^[14]	0.85	0.80	0.85	1.36
	基于 MSER 方法 ^[15]	0.88	0.86	0.87	1.13
	基于 MSER+SWT 的方法 ^[16]	0.87	0.82	0.84	1.58
实验 2	基于 MSER+NMS 的方法	0.89	0.88	0.89	1.36
实验 3	基于 MSER+NMS+形态学腐蚀的方法	0.93	0.91	0.92	1.48

从图 10、图 11、图 12 可知,基于 MSER+NMS+形态学腐蚀的方法对于两个偏旁组合的甲骨文合体字检测效果最佳,对于 3 个偏旁组成的甲骨文合体字存在部分检测框过大的现象,对于 4 个偏旁组成的甲骨文合体字检测效果较差,根据甲骨文专家提供的领域知识可知,在甲骨文合体字中,90% 甲骨文合体字由两个单偏旁组成,因此本文 MSER+NMS+形态学腐蚀的方法检测准确率高.



图 10 两个单偏旁甲骨文字符检测结果图

Fig. 10 The detection results of oracle characters with two oracle radicals



图 11 3 个单偏旁甲骨文字符检测结果图

Fig. 11 The detection results of oracle characters with three oracle radicals

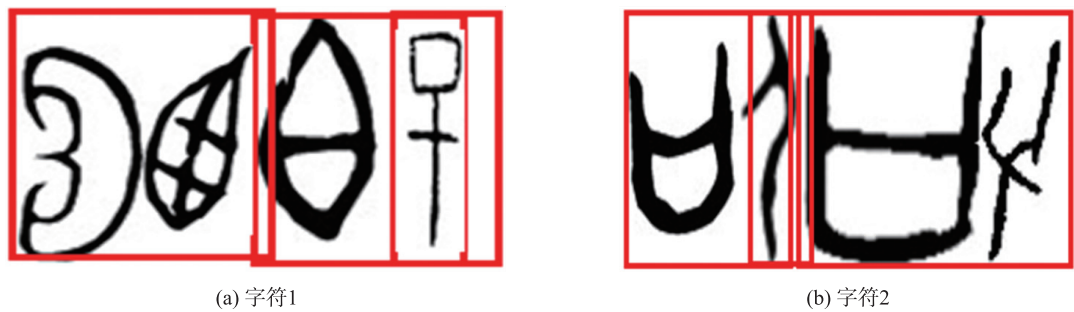


图 12 4 个单偏旁甲骨文字符检测结果图

Fig. 12 The detection results of oracle characters with four oracle radicals

3.4 甲骨文单偏旁识别对比实验分析

对于甲骨文单偏旁识别,为了对比不同网络结构对甲骨文单偏旁分类性能的影响,本文设置了 3 组实验.

实验 1:为了比较参数量对实验结果的影响,我们比较了不同结构网络模型对 Oracle 单偏旁识别的影响.

实验 2:为了验证批处理归一化层对神经网络的影响,进行了两类训练测试:(1)训练 LeNet 网络;(2)训练包括批量归一化层的 LeNet 网络.

实验 3:为了验证添加 Dropout 层和 L2 正则化的有效性,进行了 3 类训练测试:(1)训练 BN-LeNet 网络;(2)训练仅包括 dropout 层的 LeNet 网络;(3)训练仅包括 L2 正则化训练 LeNet 网络.

第 1 组实验分析:针对不同网络模型对实验结果的影响,我们比较了 5 种网络模型. 从表 2 可以看出,相比其他模型, LeNet 达到更好的精度,因此本文选择了 LeNet 网络,如图 13(b) 所示,当用相同的甲骨文单偏旁的数据集训练 LeNet 和 AlexNet 网络, AlexNet 模型在训练第 7 个 epoch 时,训练精度为 1,而测试集精度才仅仅达到 0.39,加之图 13(a) 所示的 AlexNet 模型较高的损失值,所以模型存在严重的过度拟合现象,这说明深度神经网络不能有效地提取特征信息来表征不同类别的甲骨文单偏旁之间的差异,反而参数较少的神经网络可以学习,如图 14(a) 和(b) 所示.

表 2 不同训练方法的分类准确率结果对比

Table 2 Comparison of classification accuracy results on different structures

	Model	Accuracy
实验 1	AlexNet model	0.64
	Vgg-16 model	0.78
	GoogleNet model	0.76
	ResNet model	0.74
	LeNet model	0.79
实验 2	LeNet+Batch normalization layer model	0.89
实验 3	LeNet+Dropout layer model	0.94
	LeNet+L2 regularization model	0.93
	BN-LeNet model	0.96

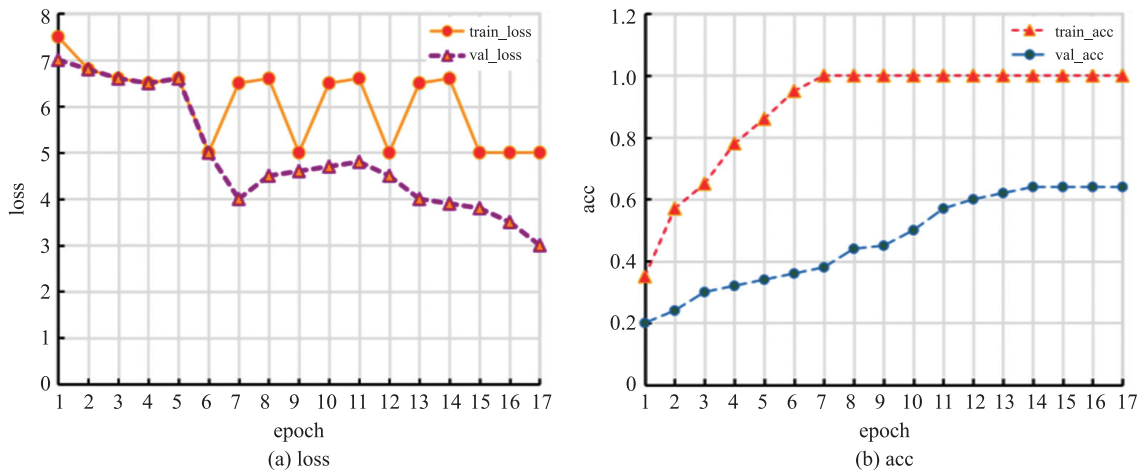


图 13 AlexNet 模型的 loss-acc 图

Fig. 13 The loss-acc diagram of the AlexNet model

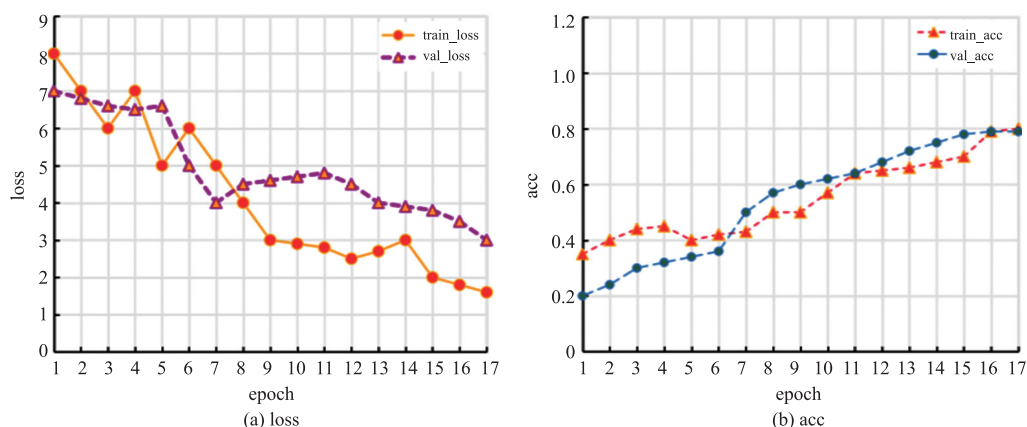


图 14 LeNet 模型的 loss-acc 图

Fig. 14 The loss-acc diagram of the LeNet model

第 2 组实验分析:如图 15(b)所示,当 LeNet 添加归一化层时,训练精度高于 LeNet,图 15(a)展示了 LeNet 添加归一化层的 loss-acc 图,我们可以看到 loss 下降速度和 acc 上升速度比单一 LeNet 网络快,这表明批归一化层的引入不仅可以加快网络的收敛速度,而且还可以提高模型的精度。

第 3 组实验分析:为了验证添加 dropout 层和 L2 正则化策略对甲骨文单偏旁分类性能的影响. 如图 16(b)所示,BN-LeNet 网络在测试集上的分类精度高达 0.96,从图 16(a)可以看出,损失值也较低,表明在甲骨文单偏旁数据集上,BN-LeNet 网络充分学习到甲骨文单偏旁特征,并且模型在训练集和测试集都表现良好,分类精度较高。

表 3 展示了部分甲骨文单偏旁手写图的预测结果。

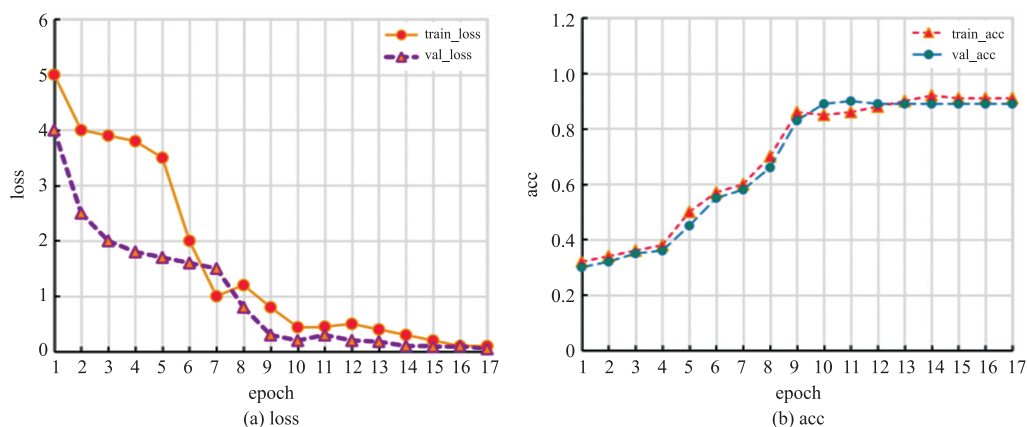


图 15 LeNet+Batch normalization layer 模型 loss-acc 图

Fig. 15 The loss-acc diagram of the LeNet with batch normalization layer

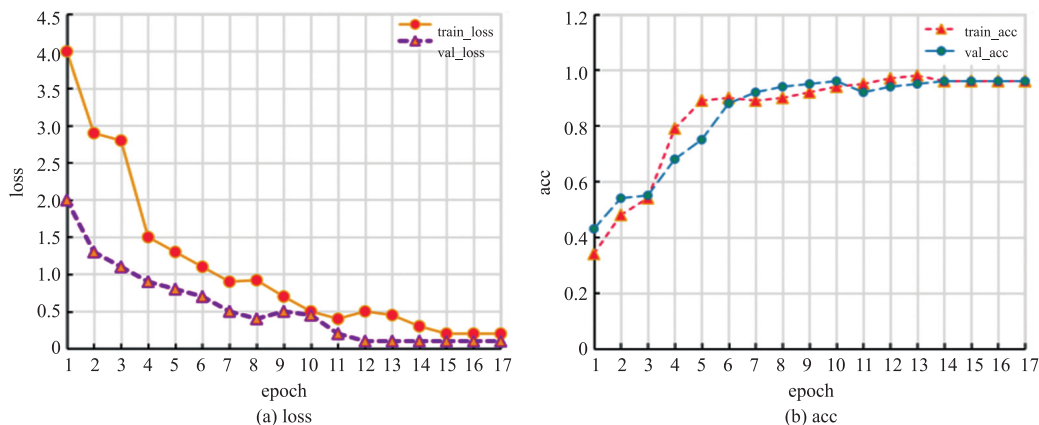


图 16 BN-LeNet 模型的 loss-acc 图

Fig. 16 The loss-acc diagram of the BN-LeNet network

表 3 甲骨文单偏旁预测结果示例(部分)

Table 3 Example of Oracle Single radical Prediction Results(Partial)

甲骨文单偏旁 手写图	数据 预处理	真实 标签	预测 标签	概率	预测 结果	甲骨文单偏旁 手写图	数据 预处理	真实 标签	预测 标签	概率
		0013	13	80.07%				0001	1	75.06%
		0004	4	73.43%				0007	7	94.49%
		0006	6	54.58%				0003	3	84.04%

3.5 甲骨文合体字识别实验分析

为了对比不同训练方法和不同网络结构对甲骨文合体字识别性能的影响,本文设置了 2 组实验.

实验 1:为了对比网络中待学习参数数量的多少对实验结果的影响,本文采用甲骨文合体字数据集对 2 种不同深度的神经网络进行训练:(1)对 LeNet 网络进行训练,待训练的参数较少,为简单的网络结构代表;(2)对 Vgg-16 网络进行训练,待训练参数相对较多,为较深的网络结构代表.

实验 2:为了验证迁移学习和微调模型的有效性,本文对甲骨文合体字数据集进 2 种不同方式的训练:(1)用在离线手写汉字 HCL2000 数据集上的预训练的 Vgg-16 网络对甲骨文合体字进行实验;(2)用微调的 OraNet 模型对甲骨文合体字进行实验,表 4 为实验结果对比.

表 4 不同训练方法的分类准确率结果对比

Table 4 Comparison of classification accuracy results on different structures

	Method	Accuracy
实验 1	LeNet model	0.82
	Vgg-16 model	0.86
实验 2	Pre-training Vgg-16 model	0.96
	OraNet model	0.98

第 1 组实验分析:针对参数数量对实验结果的影响,采用不同深度的网络做了 2 组对比实验. 从图 17(b)和图 18(b)可以看出,LeNet 网络和 Vgg-16 网络在相同的甲骨文合体字数据集上训练时,Vgg-16 模型在测试集上的分类准确率明显高于 LeNet 模型 的分类准确率,对比图 17(a)和图 18(a),Vgg-16 模型损失值较低,因此在甲骨文合体字识别时,选用较深的网络,分类准确率高.

第 2 组实验分析:如图 19 所示,预训练模型比从头训练的模型精度高,并且本文提出的微调预训练模型 OraNet 在甲骨文合体字数据集上表现更好,从图 19(a)、(b)图可看出,OraNet 模型在测试集上的分类准确率可达 0.98,说明 OraNet 模型在脱机手写汉字 HCL2000 数据集上学习到了一些有用特征,这些特征有助于对甲骨文合体字图像的分类.

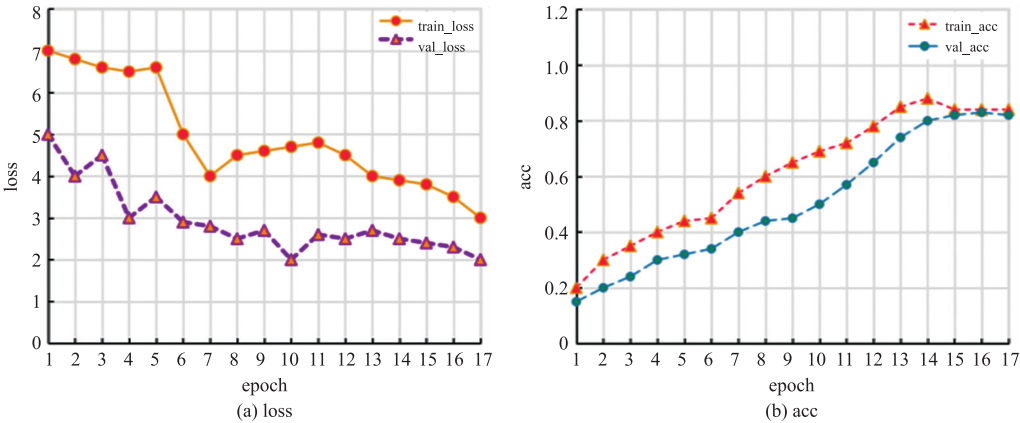


图 17 LeNet 模型的 loss-acc 图

Fig. 17 The loss-acc diagram of the LeNet network

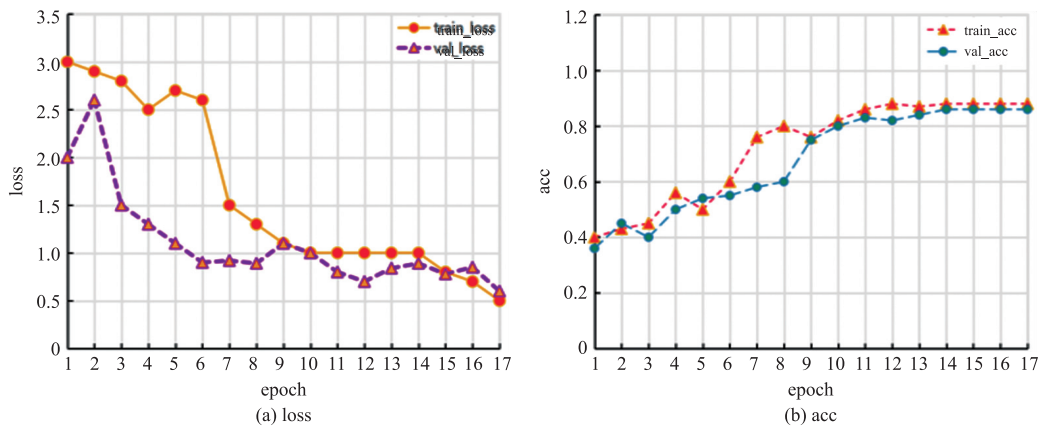


图 18 Vgg-16 模型的 loss-acc 图
Fig. 18 The loss-acc diagram of the Vgg-16 network

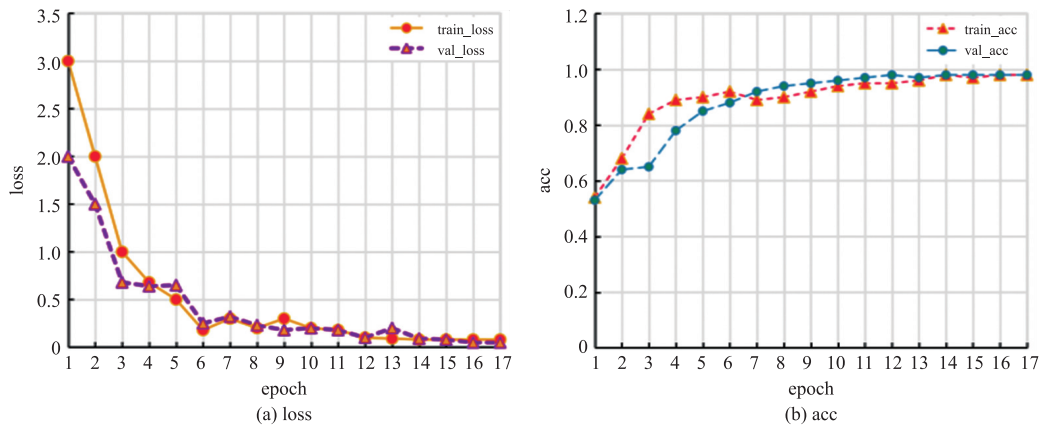
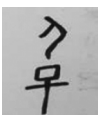
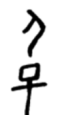




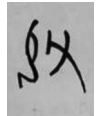







图 19 OraNet 模型的 loss-acc 图
Fig. 19 The loss-acc diagram of the OraNet network

表 5 展示了部分甲骨文合体字识别结果,本文约定甲骨文结构标记:u_x 代表该偏旁 x 在合体字上方,d_x 代表该偏旁 x 在合体字下方,l_x 代表该偏旁 x 在合体字左方,r_x 代表该偏旁 x 在合体字右方,如表 5 中第一行合体字标签为 u_7_1__d_2_0,它代表着该合体字是上下结构,7_1 偏旁位于该合体字上方,2_0 偏旁位于该合体字下方.

表 5 甲骨文合体字预测结果示例
Table 5 Example of Oracle combined character Prediction Results(Partial)

甲骨文合体字手写图	数据预处理	真实标签	预测标签	概率	结构标记	甲骨文合体字手写图	数据预处理	真实标签	预测标签	概率
		u_7_1__ d_2_0	u_7_1__ d_2_0	65.43%	上下结构 7_1 偏旁在上方 2_0 偏旁在下方			u_7_1__ d_1_1	u_7_1__ d_1_1	77.43%
		l_7_1__ r_10_1	l_7_1__ r_10_1	82.70%	左右结构 7_1 偏旁在左方 10_1 偏旁在右方			l_10_1__ r_12_1	l_10_1__ r_12_1	87.53%
		l_7_1__ r_7_1	l_7_1__ r_7_1	54.78%	左右结构 7_1 偏旁在左方 7_1 偏旁在右方			l_7_1__ r_7_1	l_7_1__ r_7_1	54.78%

4 结论

本文建立了甲骨文偏旁数据集和甲骨文合体字数据集,接着提出两种以甲骨文偏旁为构件的卷积

神经网络检测识别方法. 不仅可以对研究甲骨文相关工作者提供帮助,简化他们的数据处理相关工作,节省专家考释甲骨文的时间,同时也为计算机介入甲骨文考释提供了新思路——从偏旁的角度来实现甲骨文考释,当有新的甲骨文字出土时,便可以利用本文提出的算法来识别该新字,那将是非常有影响的成果,因此在后续的研究工作中,我们将运用该方案尝试其他少样本文字的识别.

致谢:此次实验的数据获取得到了首都师范大学甲骨文研究专家的支持,在此表示衷心的感谢!

[参考文献]

- [1] 王宇信. 政府推动下的甲骨文研究全面深入发展与弘扬阶段的时代特色[J]. 殷都学刊, 2019, 40(1): 1-6.
- [2] 国家文物局. 迎接甲骨文发现 120 周年研讨会在安阳召开[J]. 遗产与保护研究, 2019, 4(1): 80-80.
- [3] 顾绍通. 基于拓扑配准的甲骨文字形识别方法[J]. 计算机与数字工程, 2016, 44(10): 2001-2006.
- [4] 周新伦, 李锋, 华星城, 等. 甲骨文计算机识别方法研究[J]. 复旦学报(自然科学版), 1996(5): 481-486.
- [5] 李锋, 周新伦. 甲骨文自动识别的图论方法[J]. 电子科学学刊, 1996(S1): 41-47.
- [6] 栗青生, 吴琴霞, 王蕾. 基于甲骨文字形动态描述库的甲骨文输入方法[J]. 中文信息学报, 2012, 26(4): 28-33.
- [7] 张鲁宁, 左信, 刘建伟. 零样本学习研究进展[J]. 自动化学报, 2020, 46(1): 1-23.
- [8] 郭心悦, 胡沁涵, 刘纯平, 等. 基于迁移学习和批归一化的菜肴图像识别方法[J]. 计算机应用与软件, 2021, 38(3): 124-133.
- [9] ZHANG H G. HCL2000-A large-scale handwritten chinese character database for handwritten character recognition[J]. Icdar barcelona Spain, 2009: 286-290.
- [10] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of machine learning research, 2014, 15(1): 1929-1958.
- [11] 姚金良, 翁璐斌, 王小华. 一种基于连通分量的文本区域定位方法[J]. 模式识别与人工智能, 2012, 25(2): 325-331.
- [12] 冯国进, 顾国华, 郑瑞红. 基于自适应投影方法的快速车牌定位[J]. 红外与激光工程, 2003, 32(3): 285-287.
- [13] 陈善雄, 韩旭, 林小渝, 等. 基于 MSER 和 CNN 的彝文古籍文献的字符检测方法[J]. 华南理工大学学报(自然科学版), 2020, 48(6): 123-133.
- [14] JAMALUDDIN F N, AHMAD S A, NOOR S, et al. Performance of DWT and SWT in muscle fatigue detection[C]//Biomedical engineering & sciences. New York: IEEE, 2016.
- [15] WANG Z J, ZHAO Y Q, ZHAO C L. Improved MSER pedestrian detection algorithm based on TOF camera[J]. Journal of physics: conference series, 2020, 1576(1): 012013.
- [16] GUAN L, CHU J. Natural scene text detection based on SWT, MSER and candidate classification[C]//2017 2nd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2017.

[责任编辑:陆炳新]