分类号 TP391　　学校代码 10541
密　级 公开　　学　号 20213609

湖南中醫藥大學

HUNAN UNIVERSITY OF CHINESE MEDICINE

硕士学位论文

基于深度学习的中医古籍缺失文本修复和手写文字识别研究

研究生姓名：盛威
专业名称：电子信息
研究方向：深度学习
申请学位类型：专业学位
导师姓名、职称：刘伟　教授/邹北骥　教授

中国·湖南·长沙
二零二四年五月

中文摘要

目的：运用人工智能技术，特别是近年来发展迅速的深度学习技术，从中医古籍文本语料库的构建、中医古籍缺失文本的预测、中医古籍手写文字识别等方面开展研究工作，深入挖掘和萃取中医古籍的精华，凝练和提取中医古籍中所蕴含的经验和智慧，为中医古籍的研究者提供高质量的文本语料库，降低中医古籍修复专家修复古医书的难度，提升研究人员阅读古医书的效率。

方法：（1）基于构建的高质量中医古籍文本语料库及深度学习语言模型，构建ACMBBERT模型，同时对中医古籍文本语料库进行划分，在划分后的语料库上分别训练和测试N-gram模型、LSTM模型、BiLSTM模型、BERT模型，验证ACMBBERT模型在缺失文本预测任务上的优势，并将ACMBBERT模型运用到文本修复场景中。（2）使用BAGAN-GP生成对抗网络结合传统数据增强方法对原始数据集进行数据增强，通过ResNet构建分类模型验证该数据生成方式的可行性，再结合迁移学习导入预训练权重提升模型的泛化能力和收敛速度。（3）基于BAGAN-GP生成对抗网络结合传统数据增强方法构建的平衡单字数据集和CRNN算法，随机拼接成不同长度的长文本数据集，训练马王堆简帛长文本识别模型，再基于PyCorrector和马王堆医书文本语料库对ACMBBERT模型进行微调，构建马王堆医书文本纠错模型。

结果：（1）在中医古籍缺失文本的预测中，BiLSTM模型优于LSTM模型，LSTM模型明显优于N-gram模型，而ACMBBERT模型效果最优，将ACMBBERT模型运用到真实文本修复场景中，ACMBBERT-S模型达到了63.36%的hit@1，82.57%的hit@5和hit@10，ACMBBERT-MWD模型达到了37.58%的hit@1，51.42%的hit@5和57.44%的hit@10。（2）ResNet网络在扩充后的平衡数据集上训练的模型识别准确率达98.589%，比原始数据集训练的模型准确率提高了9.181%。迁移学习对比实验中，EfficientNetV2 small表现最优，准确率达99.048%。（3）基于生成数据集训练的马王堆简帛长文本识别模型在生成数据集上识别准确率较高，达86.73%，而在真实的马王堆医书简帛图像上准确率不高，仅有63%。结合马王堆医书文本纠错模型，马王堆医书简帛长文本识别模型的识别准确率得到了较大提升，达到了91%。

结论：深度学习技术应用于中医古籍缺失文本修复中具有较好的效果，能够为中医古籍修复提供帮助；BAGAN-GP生成模型与传统数据增强方法结合的方式能够很好地适用于马王堆简帛手写文字数据集的扩充。ResNet图像分类网络在平衡数据集上可以取得更高的识别准确率。结合迁移学习方式，导入预训练权重，模型的训练可以更快的收敛，准确率也相应提升；基于单字图像数据集，随机生成长文本序列数据集的方法有效地提升了文本识别任务中数据集构建的效率。基于语言模型对文本识别模型的预测进行纠错，能够较大地提升文本识别模型的准确率。

关键词：中医古籍；深度学习；文本预测；数据增强；图像识别；文本纠错

ABSTRACT

Objective: Using artificial intelligence technology, especially the rapid development of deep learning technology in recent years, the research work is carried out from the construction of the text corpus of ancient Chinese medicine books, the prediction of missing texts of ancient Chinese medicine books, and the recognition of handwritten text of ancient Chinese medicine books, so as to deeply excavate and extract the essence of ancient Chinese medicine books, condense and extract the experience and wisdom contained in ancient Chinese medicine books, provide high-quality text corpora for researchers of ancient Chinese medicine books, and reduce the difficulty of restoration experts and researchers of ancient Chinese medicine books.

Methods: (1) Based on the constructed high-quality text corpus of ancient Chinese medicine texts and deep learning language model, the ACMBBERT model was constructed, and the text corpus of ancient Chinese medicine texts was divided, N-gram model, LSTM model, BiLSTM model and BERT model were trained and tested on the divided corpus respectively, so as to verify the advantages of the ACMBBERT model in the missing text prediction task, and apply the ACMBBERT model to the text repair scenario. (2) The BAGAN-GP generative adversarial network combined with the traditional data augmentation method was used to augment the data of the original dataset, and the feasibility of the data generation method was verified by constructing a classification model through ResNet, and then the transfer learning was combined with the introduction of pre-trained weights to improve the generalization ability and convergence speed of the model. (3) Based on the balanced single-word dataset and CRNN algorithm constructed by the BAGAN-GP generative adversarial network combined with the traditional data augmentation method, the long text dataset of different lengths was randomly spliced into a long text dataset of different lengths, and the Mawangdui simple silk long text recognition model was trained, and then the ACMBBERT model was fine-tuned based on PyCorrector and the Mawangdui medical book text corpus to construct the Mawangdui medical book text error correction model.

Results: (1) In the prediction of missing texts in ancient TCM books, the BiLSTM model was better than the LSTM model, the LSTM model was significantly better than the N-gram model, and the ACMBBERT model had the best effect. In the real text restoration scene, the ACMBBERT-S model, reached 63.36% hit@1 and 82.57% hit@5, the ACMBBERT-MWD model reached 37.58% hit@1, 51.42% hit@5, 57.44% hit@10. (2) The recognition accuracy of the model trained by the ResNet network on the expanded equilibrium dataset is 98.589%, which is 9.181% higher than that of the model trained on the original dataset. In the comparison experiment, EfficientNet V2 small performed the best, with an accuracy of 99.048%. (3) The recognition accuracy of the Mawangdui simple silk long text recognition model trained based on the generated dataset is high,

reaching 86.73%, while the accuracy of the real Mawangdui medical book is not high, only 63%. Combined with the Mawangdui medical book text error correction model, the recognition accuracy of the Mawangdui medical book text recognition model has been greatly improved, reaching 91%.

Conclusion: Deep learning technology has a good effect in the restoration of missing texts of ancient Chinese medicine books, and can provide help for the restoration of ancient Chinese medicine books. The combination of BAGAN-GP generation model and traditional data augmentation methods can be well applied to the expansion of Mawangdui handwriting dataset. ResNet can achieve high recognition accuracy on the expanded balanced dataset. Combined with the transfer learning method and the introduction of pre-training weights, the training of the model can converge faster and the accuracy can be improved accordingly. Based on the single-word image dataset, the method of randomly generating long text sequence datasets effectively improves the efficiency of dataset construction in text recognition tasks. The error correction of the prediction of the text recognition model based on the language model can greatly improve the accuracy of the text recognition model.

Key words: Ancient Books of Traditional Chinese Medicine; Deep Learning; Text Prediction; Data Augmentation; Image Recognition; Text Error Correction

目　录

第一章 绪论

1.1 研究背景与意义

中医古籍是中医药学传承和发展的载体，既具有重要的学术价值，又具有相当的文物价值，中医药古籍文献中所收载的理法、方药、养生保健知识是取之不尽、用之不竭的宝库，具有极高的实用价值[1]。中医古籍反映了中医药学在历史上为中华民族的繁衍昌盛作出的伟大贡献和取得的辉煌成就，更重要的是，它所蕴藏保存下来的民族文化遗产和智慧结晶必将在新的历史时期为人类医药科学事业的发展作出不可估量的贡献[2]。

但中医古籍在长期的流传过程中，不仅遭受自然的侵蚀，也会受到人为的损坏，同时中医专业图书馆的古籍藏书可能由于没有得到很好的保存，即使经过修复，仍然存在破损的现象，而且图书馆的书库环境、保护手段都可能造成中医古籍的损坏。因此针对中医古籍进行整理和修复是十分必要的，而目前的修复工作主要是由古籍修复专家通过专业手段手动修复，但这个过程需要对古汉语和中医古籍有广泛且深入的研究，非常耗时费力。

由于中医古籍为古汉语书写，古汉语文本的整体信息化程度较低，古汉语特别是医古文与信息处理技术缺乏有机结合等问题，关于中医古籍的信息化和智能化研究鲜有报道。因此，本文将力图填补此项空白，对中医古籍的缺失文本修复和手写文字识别开展深入研究，以人工智能技术助力中医古籍的传承与创新。

运用人工智能技术，特别是近年来发展迅速的深度学习技术，从中医古籍文本语料库的构建、中医古籍的缺失文本修复、中医古籍手写文字识别等方面开展一系列研究工作，深入挖掘和萃取中医古籍的精华，凝练和提取中医古籍中所蕴含的经验和智慧，可以为中医古籍的研究者提供高质量的文本语料库，降低中医古籍修复专家和研究人员修复古医书的难度。

总之，本文将围绕中医古籍开展数字化、信息化、智能化和系统化研究，对中医古籍文本语料库构建、缺失文本修复和中医古籍手写文字识别等开展一系列研究，推动中医文化传承和中医复兴。

1.2 国内外研究现状

1.2.1 古汉语语料库构建研究现状

古汉语信息处理是当前中文信息处理领域一个重要研究方向，对古籍文献进行数字化和信息化的首要任务就是构建古汉语语料库。文献[3]对先秦文献的语料库构建进行了初步的探究，阐述了古文献语料库构建过程中涉及的古汉语分词以及词性标注等基本知识。在构建古籍语料库时需要制定对应的分词规范和标注集，南京师范大学语言科技研究所制定了先秦汉语词类标注基本集并构建了先秦典籍人工语料库[4]。文献[5]以《淮南子》为例介绍

了上古汉语分词及词性标注语料库及其构建过程。文献[6]运用条件随机场方法对先秦汉语的分词和标注一体化工作开展研究。这些研究工作都没有考虑到中医古籍的特点，因此，在构建中医古籍语料库时，需要充分融入中医古籍内在特点对相应的方法进行合理的优化和改进。文献[7]对构建中医药古文献语料库的意义和思路进行了阐述，并且对语料分类问题进行了详细的介绍。文献[8]对中医古籍的分词规范进行了探索性的研究和总结，提出中医古籍分词规范建议，并以人工标注方式构建了一个小型的清代医籍分词语料库，但未能实现计算机自动分词，且语料库规模较小，不能推广和应用。

1.2.2 古籍缺失文本修复研究现状

由于年代久远，许多出土帛书残损非常严重，有诸多帛书碎片缺失，导致古医书文字残缺不全。如何对这些残字进行考释和修复是中医古籍研究人员面临的一个非常重要的问题。近年来，国外有学者开始尝试将深度学习技术应用于古文本的修复工作并取得了较好的效果。2019年，牛津大学与Google旗下国际顶尖的人工智能（Artificial-Intelligence，AI）企业DeepMind合作开展古文修复研究，联合开发了一款名为Pythia的AI工具，利用深度学习神经网络技术来识别和预测古希腊碑文中缺少的单词或字符，取得了较好的效果[9]。2020年9月，国际顶级期刊PNAS发表了几位以色列学者运用深度学习中的循环神经网络（Recurrent Neural Networks，RNN）来恢复残缺的古巴比伦文本的一项研究成果，实验结果表明该模型在预测残缺文本时效果良好，可以作为学者进行文本恢复工作的有力工具[10]。

常用的语言模型分为以N-gram[11]模型为代表的传统统计语言模型和以NNLM[12](Neural Net Language Model)、长短期记忆（Long Short-Term Memory，LSTM）模型[13]为代表的神经网络语言模型，以及后续由神经网络语言模型发展来的以BERT[14]（Bidirectional Encoder Representation from Transformers）为代表的预训练语言模型，如GPT[15]（Generative Pre-Training）、RoBERTa[16]（A Robustly Optimized BERT Pre-training Approach）、XLNet[17]（Generalized autoregressive pretraining for language understanding）等一系列模型。预训练语言模型相较于传统的统计语言模型和神经网络语言模型，在大规模语料的处理上更具优势，预训练提供的模型初始化参数，可以使模型在目标任务上有更好的泛化性能和更快的收敛速度，同时在小规模数据上也能避免过拟合现象的发生。虽然预训练语言模型取得了很大成功，但是目前大多数预训练模型都是基于大量通用语料训练的，在面对特定领域文本的自然语言处理任务时，其功能的发挥容易受限。由于古代汉语在语法、语义、语用上与现代汉语存在较大差异，即使是面向中文构建的Chinese-BERT[18]，在古汉语处理上也难以达到其在中文通用语料上的性能，直接调用当下的预训练模型预测中医古籍中缺失的文本，并不能取得较好的效果。

1.2.3 古籍手写文字识别研究现状

古籍是研究历史文化，挖掘古人智慧的重要载体。针对古籍的文字识别，是古籍数字

化的一大重要研究问题。文字识别属于光学字符识别的范畴，但是古籍的文字识别又存在其特殊性。古籍中的古汉字写法，在历朝历代都存在差异性，其次存在大量的生僻字，数据有着很大的不平衡性。此外，古籍由于年代久远等问题，图像不清晰，甚至有破损也是一大问题。

在传统的文本识别方法中，文本识别任务分为3个步骤，即图像预处理、字符分割和字符识别。传统方法需要对特定场景进行建模，一旦场景变化就会失效。面对复杂的文字背景和场景变动的情况，基于深度学习的方法具有更优的表现。基于深度学习的文本识别方法按照场景可以分为规则文本识别[19-26]和不规则文本识别[27-34]。规则文本识别又分为基于CTC（Conectionist Temporal Classification）的算法和基于Sequence2Sequence的算法，两种算法在规则文本识别领域都被证明是有效的。基于CTC最典型的算法是CRNN[35]（Convolutional Recurrent Neural Network），CRNN算法使用主流的卷积结构提取图像特征，常用的卷积结构有ResNet、MobileNet、VGG等，同时引入双向LSTM模型用于增强上下文建模，解决了卷积神经网络缺乏长依赖建模能力的问题。Xie等人[36]提出聚合交叉熵(ACE)损失，减轻了CTC反向传播负担，优化了每个字符的统计频率，大大提高了效果。Sequence2Sequence算法是由编码器Encoder把所有的输入序列都编码成一个统一的语义向量，然后再由解码器Decoder解码。在解码器Decoder解码的过程中，不断地将前一个时刻的输出作为后一个时刻的输入，循环解码，直到输出停止符为止。一般编码器是一个RNN，对于每个输入的词，编码器输出向量和隐藏状态，并将隐藏状态用于下一个输入的单词，循环得到语义向量；解码器是另一个RNN，它接收编码器输出向量并输出一系列字以创建转换。受到Sequence2Sequence在翻译领域的启发，Shi[37]提出了一种基于注意的编解码框架来识别文本，通过这种方式，RNN能够从训练数据中学习隐藏在字符串中的字符级语言模型。以上两个算法在规则文本上都有很不错的效果，但由于网络设计的局限性，这类方法很难解决弯曲和旋转的不规则文本识别任务。而中医古籍的文本通常是竖直排列的，相对于自然场景的文本识别任务来说是比较规则的，因此现有的文本识别方法可以应用于古籍的文本识别，同时也需要考虑古籍文本的特殊性。

1.3 主要研究内容

论文主要研究中医古籍缺失文本的修复和中医古籍手写文字的识别。通过对语言模型和文字识别技术进行研究，结合相关技术特点和发展，将深度学习技术成功应用于中医古籍的缺失文本预测和手写文字识别。主要使用循环神经网络技术，卷积神经网络技术，语言模型技术，生成对抗网络技术等。论文针对两个任务分别构建了高质量的中医古籍文本语料库和马王堆简帛手写文字图像数据集，同时使用生成对抗网络技术解决了中医古籍简帛手写文字数据不平衡问题。论文在这两个自建数据集上分别进行缺失文本预测实验和古籍手写文字识别实验，验证模型效果。同时实现了一个中医古籍缺失文本预测和古籍手写

文字识别系统。针对以上阐述，论文主要工作有：

（1）收集整理了一份83 032 043字的中医古籍语料库。该语料库收录了自秦汉以来的包括《黄帝内经》《金匮要略》《伤寒杂病论》《神农本草经》等经典中医古籍在内的608本中医古籍。

（2）构建了一份3 339类，每一类约3 000个样本的马王堆简帛手写文字平衡图像数据集。当前含有比较多的手写英文数据集，公开的手写中文数据集比较稀少，而公开的中医古籍手写文字图像数据集则更加稀少。《马王堆汉墓简帛文字全编》对马王堆所有简帛文字原始资料作了通盘的清理，秉持“有字必录”原则，收录了马王堆汉墓简帛中出现的3 339类文字。本文以《马王堆汉墓简帛文字全编》为基础，使用目标检测和分割算法，自动分割出《马王堆汉墓简帛文字全编》中所列出的单字图像，然后人工进行分类，形成原始单字图像数据集。而原始的单字图像数据集存在极大的数据不平衡性，因此论文又使用图像增强技术和生成对抗网络技术对原始图像数据集进行扩充，最终形成了平衡的马王堆简帛手写文字平衡图像数据集。

（3）将语言模型技术应用于中医古籍缺失文本的预测。基于中医古籍语料库，构建基于中医古籍的预训练BERT模型（Ancient Chinese Medical Book Bidirectional Encoder Representation from Transformers，ACMBBERT），同时分别训练N-gram模型、LSTM模型、BiLSTM模型，与ACMBBERT模型进行对比实验。实验结果证明BiLSTM模型优于LSTM模型，LSTM模型明显优于N-gram模型，而ACMBBERT模型效果最优，将ACMBBERT模型运用到《黄帝内经》的修复中，达到了63.36%的hit@1，82.57%的hit@5。即在表现效果最优的ACMBBERT输出的前10条预测中，可以直接获得正确的缺失内容的几率为60.79%，出现正确预测的几率为82.57%。此外，目前已有如GuwenBERT[38]、SikuBERT[39]的预训练语言模型可以为古文分词、古文命令实体识别等下游任务提供强大的文本特征提取能力，而在中医古籍自然语言处理领域，本论文训练的ACMBBERT模型还可进一步应用于中医古籍的分词、机器翻译、命令实体识别等下游任务中。

（4）提出了一种结合语言模型纠错的马王堆医书长文本识别模型。基于马王堆简帛平衡单字图像数据集，从每一类中随机抽取一张图片，拼接成随机长度的文本列，构建长文本数据集，作为长文本识别模型的训练数据。如此，可以生成大量长度不一的训练数据，无需从马王堆简帛原始图像中截取文本列，再不断地人工查阅《马王堆汉墓简帛文字全编》添加标签，极大提升了数据集构建的效率。基于生成的长文本数据集和CRNN算法，训练马王堆简帛长文本识别模型。该模型在生成数据集上识别准确率较高，达86.73%，而在真实的马王堆医书简帛图像上准确率不高，仅有63%。结合在ACMBBERT模型和马王堆医书文本语料库基础上微调来的马王堆医书文本纠错模型，马王堆医书简帛长文本识别模型的识别准确率得到了较大提升，达到了91%。同时，对于《天回医简》中截取的包含25幅单字图像和25幅长文本序列图像的测试集，马王堆简帛长文本识别模型能够准确地识别出其

中19幅单字图像，虽然仅能完全准确地识别出4幅长文本序列图像，但是只有一个字识别错误的有12幅，同时25幅长文本序列图像共计134个字，模型准确识别的总字数为87，在长文本图像的字符维度上达到了60.44%的准确率。

（5）设计和实现了一个中医古籍缺失文本预测和手写文字识别系统。系统支持用户输入隐去了一个或多个字的中医文本，并输出缺失文本预测模型的预测。也支持用户输入马王堆简帛中的某一张单字图像或某一列图像，快速获取图像中文字内容，从而降低马王堆简帛阅读门槛。

1.4 组织结构安排

第一章：绪论。首先介绍了论文相关研究的背景与意义以及古汉语语料库构建、古籍缺失文本修复、古籍手写文字识别研究现状的研究现状，并对论文主要研究内容和组织结构安排进行了介绍。

第二章：相关技术与语料库构建。首先介绍了在中医古籍缺失文本预测和中医古籍简帛手写文字识别所需的理论和技术。然后介绍了中医古籍文本语料和简帛手写图像数据集的构建过程。着重介绍了如何使用生成对抗网络及传统数据增强方法解决原始简帛单字图像数据集的数据不平衡问题。

第三章：中医古籍缺失文本预测研究。首先介绍了ACMBBERT模型构建研究的总体流程，然后介绍了中医古籍缺失文本预测研究的评价指标和实验流程。在自建的中医古籍语料库上训练了各语言模型，验证了ACMBBERT模型在缺失文本预测任务上的优势。将ACMBBERT模型运用到中医古籍文本修复场景中，达到了训练一个中医古籍文本修复模型的预期目标。

第四章：中医古籍简帛手写文字识别研究。首先使用BAGAN-GP生成对抗网络与传统数据增强方法结合的方式构建平衡马王堆简帛单字图像数据集，并通过ResNet构建分类模型验证该数据生成方式的可行性，再结合迁移学习导入预训练权重提升模型的泛化能力和收敛速度。其次介绍了马王堆简帛手写长文本识别模型的构建流程，基于构建的平衡单字数据集和CRNN算法，随机生成了大量的文本序列图像数据，训练文本识别模型，再基于马王堆医书文本语料库对先前构建的ACMBBERT模型进行微调，构建马王堆医书文本纠错模型。将马王堆医书文本纠错模型与文本识别模型进行结合，从图像和文本两个角度，提升马王堆医书简帛图像长文本识别的准确率。

第五章：中医古籍智能修复系统的设计和实现。首先介绍了Flask框架，然后对系统需求进行了分析，介绍了系统三大功能模块基本需求。接着介绍了系统总体操作流程，最后介绍了各功能模块的实现过程以及界面展示。

第六章：总结与展望。总结了本文的研究成果和研究内容，同时分析后续存在改进的工作，最后对未来的研究进行展望。

第二章 相关技术与数据集构建

2.1 语言模型

语言模型（Language Model，LM）[40]在自然语言处理（Natural Language Processing，NLP）[41]领域中是实现自然语言理解和生成的关键组件之一。语言模型通过学习语言的统计规律，能够预测文本中下一个词或短语的出现概率，从而实现对自然语言的理解和生成。它在诸如机器翻译、语音识别、文本生成、情感分析等众多NLP任务中发挥着重要作用，为自然语言处理领域的发展提供了强大的支持。

作为自然语言处理领域的核心组件，语言模型不仅能够提供词的向量表示，还能够计算词序列的联合概率。在构建语言模型时，我们依据文本中的先行词信息，预测下一个可能出现的词，从而解决文本序列数据的预测问题。通过这种方法，语言模型能够有效地捕捉语言的内在规律和结构，为NLP任务的执行提供有力的支持。

对于给定的词序列$W =< W_1W_2W_3 \dots W_N >$，语言模型通过计算给定词序列$W$的概率$P(W)$来判断词序列$W$是否能够作为一个句子。计算$P(W)$的公式见式(2-1)-(2-3)所示[42]。

$$P(W) = P(W_1, W_2, W_3, \dots, W_N) \tag{2-1}$$

$$P(W) = P(W_1)\,P(W_2|W_1)P(W_3|W_1W_2) \dots P(W_n|W_1 \dots W_{N-1}) \tag{2-2}$$

$$P(W) = \prod_{i=1}^{N} P(W_i|W_1, W_2, \dots, W_{i-1}) \tag{2-3}$$

其中$P(W_1, W_2, W_3, \dots, W_N)$表示词$W_1$到$W_N$同时存在发生的概率，$P(W_i)$表示句子的第$i$个词是词$W_i$的概率，$P(W_n|W_1 \dots W_{N-1})$的表达式为$\frac{P(W_N)}{P(W_i|W_1,W_2,\dots,W_{N-1})}$。计算概率同时还需要满足两个约束条件，见式(2-4)所示。

$$\begin{cases} P(W_i|W_1, W_2, \dots, W_{i-1}) > 0 \\ \sum_{w} P(W_i|W_1, W_2, \dots, W_{i-1}) = 1 \end{cases} \tag{2-4}$$

语言模型的发展经历了三个阶段，传统概率语言模型阶段，循环神经网络语言模型阶段和近年来的预训练语言模型阶段。传统概率语言模型阶段，以N-gram为带代表的语言模型使用统计方法来描述自然语言的内在规律，这种基于概率模型的语言模型简单而有效，在早期得到广泛应用。然而，基于N-gram模型的语言模型，会因为数据稀疏和长距离信息的问题，在性能上受到限制。而随着神经网络技术的飞速发展，循环神经网络技术逐渐成为主流的语言模型技术，其通过其循环结构，可以记忆长距离的词信息，解决了N-gram语言模型的数据稀疏问题，但是会导致梯度消失和梯度爆炸的问题[42]。近年来，预训练语言模型取得了很大成功，其使用自监督学习从大规模文本数据学习普适性的语言表征，然后将学到的知识迁移到下游任务。这些模型能为下游任务提供强大的特征提取能力，从而可

避免从头开始训练下游任务。

2.1.1 N-gram语言模型

N-gram语言模型是以Markov[43]理论为假设前提，即句子中第N个词出现的概率只与前N-1个词有关，与其他词无关。因此整个句子的概率等于每个词出现的条件概率的连续相乘，则表示词W_i有式(2-5)：

$$P(W_i|W_1^{i-1}) = P_{NG}(W_i|W_{i-N+1}^{i-1}) \quad (2-5)$$

$W_1^{i-1} =< W_1, W_2, W_3, \ldots, W_{i-1} >$，满足上述条件的就是N-gram语言模型。其中N越大，模型性能就会越准确，同时复杂度也会越高，计算量会越大。因此在实际场景中，研究者通常取N为2或3。由此可知，第N个词的生成概率是由前面的N-1个词共同决定，这就导致它过于依赖训练语料的丰富程度，否则就很容易出现数据稀疏问题，并且计算复杂度会随着N的增加而产生指数级的增长。

2.1.2 循环神经网络语言模型

RNNLM结构图如图2.1所示，在每个时间步t，RNNLM可表示为式(2-6)-(2-8)[44]：

$$x_t = [\omega_t^T; S_{t-1}^T]^T \quad (2-6)$$

$$s_t = f(Ux_t + b) \quad (2-7)$$

$$y_t = g(Vs_t + d) \quad (2-8)$$

其中U、V是权值参数，d、b分别是输出层和状态层的偏置参数。$f(.)$是 Sigmoid激活函数，$g(.)$则是Softmax函数。RNNLM的主要优势在于可以利用基于时间的反向传播算法来训文本数据。由于文本数据具有时序性，因此RNNLM 的性能要明显优于N-gram模型。

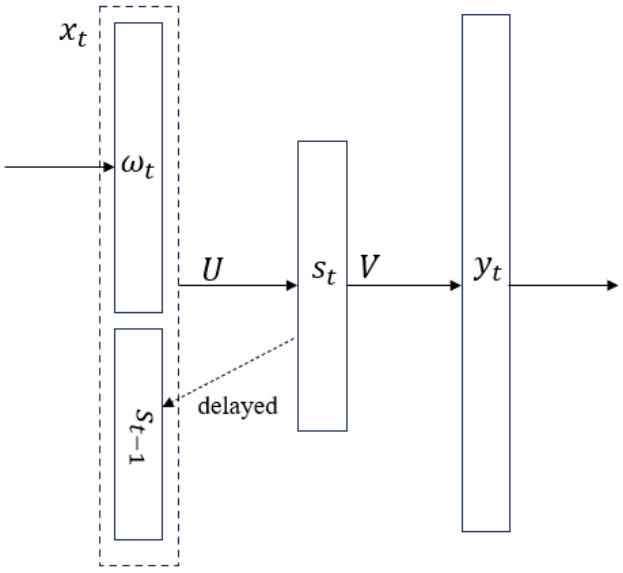

图2.1 循环神经网络语言模型结构图

相比于N-gram语言模型，循环神经网络语言模型具备了记忆功能，可以捕捉序列中的

长期依赖关系，在预测时可以做到考虑上下文信息，但是在处理长序列时，可能会发生梯度消失或梯度爆炸的问题。为了解决这些问题，LSTM语言模型通过引入门控机制有效缓解了梯度消失的问题，而梯度裁剪方法，在每次更新权重之前，检查梯度的范数，在其超过某个阈值时将其缩放到一个较小的值，有效防止了梯度爆炸的问题。近年来，注意力机制被引入到语言模型任务上，其允许模型对输入序列的不同位置分配不同的权重，以便在处理每个序列元素时专注于最相关的部分，使得模型的性能大幅提升。

2.1.3 预训练语言模型

当前研究显示，通过在庞大的无标注语料库上进行预训练，能够显著提升NLP任务的模型性能。这一发现开启了迁移学习的新篇章，目前主流的迁移学习方式主要有两种：特征提取和微调。特征提取方法，以ELMo（Embeddings from Language Models）[45]等模型为代表，其特点在于冻结预训练参数。这意味着在迁移过程中，预训练模型的结构和权重保持不变，仅用于提取特征，供下游任务使用。这种方式的优点在于保持了预训练模型的原始特性，但可能需要更为复杂的任务架构来适应。相比之下，微调方法则以BERT等模型为代表，它允许在迁移过程中动态地调整参数。这意味着在针对具体下游任务时，可以对预训练模型的参数进行微调，以更好地适应任务需求。这种方法通常更为灵活，能够更好地捕捉任务特定的信息，因此在综合性能上往往优于特征提取方法。目前，以GPT和BERT为代表的预训练语言模型，基于大规模文本训练得出，已成为主流的文本表示模型。

BERT是一种基于Transformer[46]结构的预训练语言模型，由Google在2018年提出，是自然语言处理领域的重要突破之一。如图2.2所示，BERT模型由多层的Transformer编码器堆叠而成，每一层都包含自注意力机制和前馈神经网络。这种深层结构使得BERT能够捕捉从浅层语法特征到深层语义特征的不同级别的语言信息。相比传统的单向语言模型，BERT在各种自然语言处理任务中取得了显著的性能提升。BERT的核心思想是通过大规模无监督预训练来学习通用的语言表示，然后在特定任务上进行微调。BERT模型的预训练包括两个阶段：掩码语言模型（Masked Language Model，MLM）和下一个句子预测（Next Sentence Prediction，NSP）分别捕获词级和句子级表示。MLM任务在输入句子中随机MASK掉一部分词汇，然后让模型预测这些被MASK的词汇是什么，从而让模型学习上下文相关的词汇表示，NSP任务给定两个句子A和B，让模型判断B是否是A的下一句，从而让模型学习句子之间的关系。通过这两个预训练任务，BERT模型能够在大规模语料库上学习到通用的语言表示，为后续的特定任务提供了强大的基础。在预训练完成后，BERT模型可以通过微调的方式在特定任务上进行训练。微调过程将BERT模型与特定任务的输出层相连接，并进行端到端的训练，使得BERT模型能够针对特定任务进行优化，进一步提高其在该任务上的性

能。

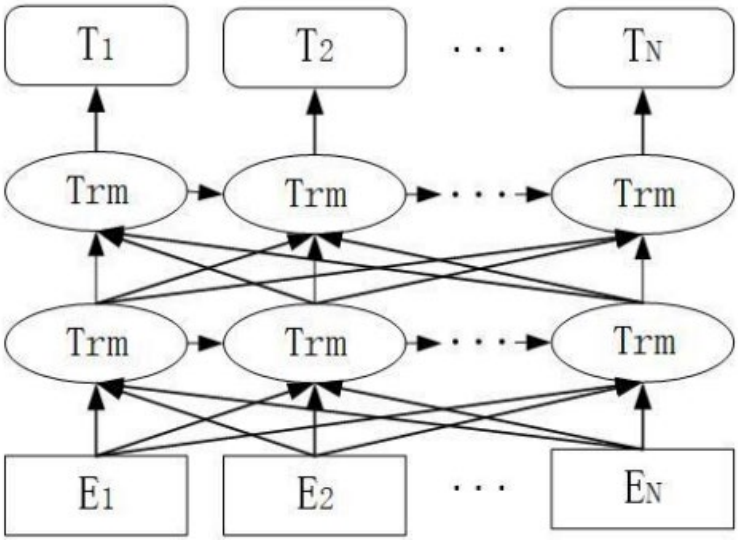

图2.2 BERT模型架构图

传统的语言模型受限于单向建模，即只能从左至右或从右至左对文本进行编码，这导致了在同时尝试顺序和逆序建模时，文本信息的泄露问题，使得模型难以深入挖掘深层次的语义信息。尽管ELMo模型通过逆序方法尝试从不同方向进行训练，但其独立的训练方法并未根本解决单一训练方向所带来的问题。实际上，ELMo在预测一个词时，仍未能同时利用到该词周围的上下文信息。为了真正实现双向文本建模，BERT引入了MLM任务。在MLM任务中，BERT会将输入文本中约15%的单词替换为掩码标记[MASK]，随后利用深层Transformer模型将这些[MASK]标记还原为原始单词。这种设计既巧妙地避免了双向语言模型的信息泄露问题，又使得模型能够同时利用上下文信息进行预测。然而，直接引入[MASK]标记会导致预训练阶段与下游微调任务之间存在不一致性，因为在实际应用中，文本中并不会出现这样的掩码标记。为了解决这一问题，BERT采取了更为精细的策略：对于选取的遮掩词，80%替换为[MASK]标记，10%替换为随机词，而剩余的10%则保留原单词。这样的处理既保证了模型在预训练时能够学习到上下文信息，又使得预训练阶段与下游任务更为接近。

BERT模型的成功之处在于它首次将无监督的预训练与有监督的微调这一模式引入到更深层的双向结构中，为自然语言处理领域带来了里程碑式的提升。在NLP领域，存在大量无监督数据，通过在这些数据集上进行预训练，可以得到一个相对完善的语言模型。当处理具体的NLP任务时，通过迁移学习，并结合任务特性进行有监督的微调，可以大大节约训练模型所需的时间和资源。

2.2 生成对抗网络

生成对抗网络（Generative Adversarial Network，GAN）[47]是一种深度学习模型，主要

构成为生成器（Generator）和判别器（Discriminator）。在对抗训练中，生成器将噪声转化成样本，并将其用于生成对抗网络的训练中；而判别器则判断训练样本是真或假，若判断正确，则提升自己的性能。二者之间通过对抗训练，以互相提升，最终达到提高模型的整体性能的目的。

BAGAN[48]是一种改进的GAN模型，它通过解决传统GAN中存在的样本不平衡问题来提高生成器的性能。BAGAN基于普通的GAN，引入了自动编码器对判别器和生成器进行初始化，同时在对抗训练阶段引入平衡因子，即根据训练数据中真假比例对样本进行权重改变，使得二者的数量达到平衡，让整个网络训练更加充分，从而缓解样本不平衡问题。最终BAGAN相比其余GAN能更好的面对样本不平衡问题。

BAGAN-GP[49]在2021年被提出。其在BAGAN的基础上，引入基于中间嵌入模型的自动编码器以及梯度惩罚。BAGAN的自动编码器不能直接学习各个类的标签信息，因此不能将各个类的样本进行分散，从而使标记信息不明确，而中间嵌入模型生成标记信息的潜在向量，更好地帮助自动编码器分散各个类的样本。对于损失函数，其引入了Wasserstein距离，来解决鉴别器过强不能给生成器有意义的梯度的问题，同时引入梯度惩罚，使对抗训练更加稳定。

2.3 常见文本识别算法

2.3.1 CRNN 文本识别方法

文字识别的目标是对定位好的文字区域进行识别,主要解决的是将一串文字图片转芦苇对应字符的问题。常用的文字识别框架主要有两个，CNN+RNN+CTC 和 CNN+Seq2Seq+Attention，CRNN是目前主流的端到端的文本识别模型，可以识别比较长的文本序列，网络的整体结构如图2.3所示。CRNN包含CNN特征提取层和Bi-LSTM序列特征提取层，能够进行端到端的联和训练。同时，CRNN利用Bi-LSTM和CTC Loss学习文字的上下文关系，从而提神了文字的识别率，是的模型更加健壮。CNN+Seq2Seq+Attention相较于CRNN，主要使用Seq2Seq+Attention做解码，可以用于多对多的映射，同时利用Attention机制，提升了文字的识别率。Attention在英文识别上的效果非常好，但是在中文识别上效果不稳定且算力开销大，所以工业界还是使用CTC居多。

CRNN模型主要分为两个部分：一部分特征提取，由多个卷积层、池化和非线性层组成，另一部分为序列预测，由RNN+CTC模型组成。RNN部分主要用于学习和建模CNN中提取到的隐藏状态以及空间特征之间的关系，最后预测初步的序列结果。粗糙的预测序列中可能存在字母重复的情况，通过CTC模块对RNN的序列进行整合，可以对序列进行去重

操作。

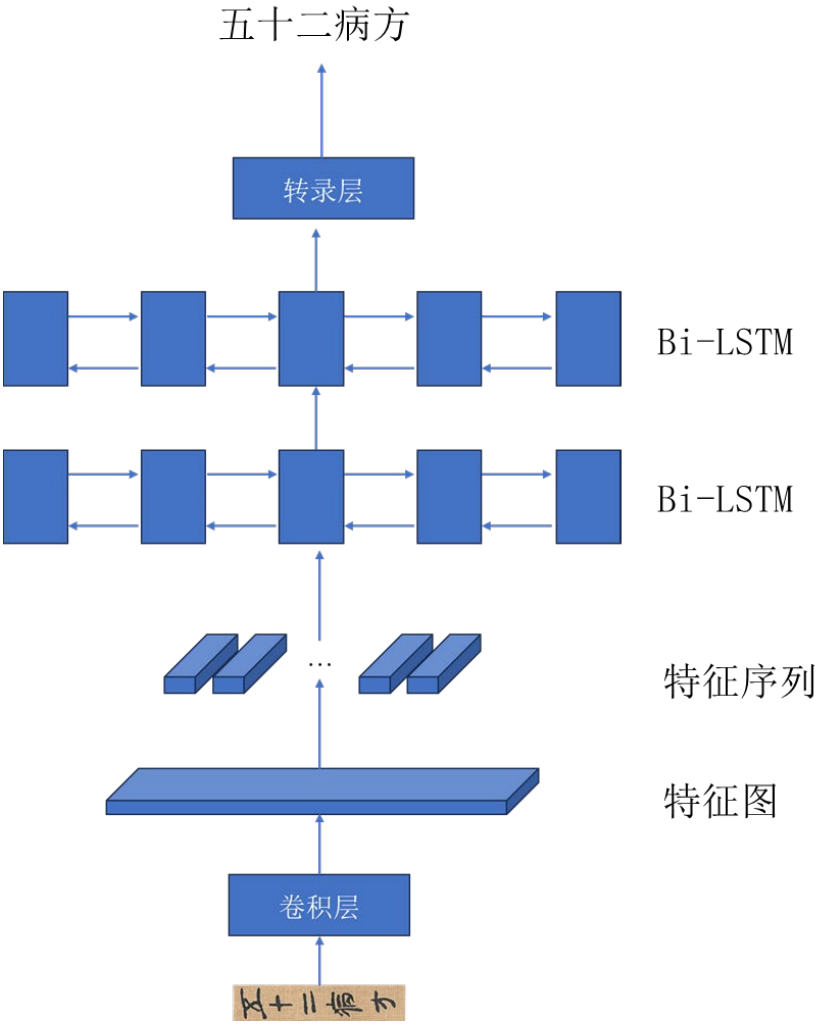

图2.3 CRNN算法结构示意图

CRNN接收灰度图或RGB彩色图作为输入，CNN作为编码器来提取与图片对应的中间层特征。经过变形后整理成T个时间步的输入送入随后的解码器RNN，从而预测出初步的序列。初步的序列经过CTC整流处理，去除冗余的字符后得到最终的预测结果。

CRNN共包含7个卷积层，2层Bi-LSTM，网络在对特征降维时最大值池化采用的窗口高度为2，这意味着每次池化高度都会减少一半，经过5次池化，高度缩减为1，宽度为原图长度的1/4。因此，序列的长度必须超过图片中单词的长度，这样才能够预测出完整的词语。

2.3.2 文本识别算法的评价指标

准确率（Accuracy）是文本识别的一种评价指标。目前准确率的计算有两种，一种是单词级别，另一种是文本行级别。文本行的准确率是计算所有要识别的文本行中有多少行被完全正确识别的比例。可以用公式（2-9）表示。

$$\text{Accuracy} = \frac{\text{Num}_{\text{correct}}}{\text{Num}} \qquad (2-9)$$

其中$\mathrm{Num}_{\mathrm{correct}}$表示识别正确的文本行的条数，Num表示所有要进行识别的文本行的总数。字符级别的准确率计算公式一致，只不过计算是按照所有字符的个数作为分母，分子是识别正确的字符个数。文本行级别的准确率要求比较严格，字符级别的准确率可以更好地评判模型的性能。

2.4 Flask 框架

Flask是一个基于Python的轻量级Web微框架，它允许开发者迅速搭建起一个功能完备的网站或Web服务。与另一著名的Python Web框架Django相比，Flask的设计更为简洁，主要适用于小型网站的开发。Django则因其丰富的内置模块，更适合大型网站的建设。Django以其一站式解决方案而著称，从设计之初就提供了模板、表单、路由以及基本的数据库管理等内建功能，这大大节省了开发者的时间。然而，Flask则采取了不同的策略。它主要依赖于Jinja2模板引擎和WSGI工具集—Werkzeug，来构建网页的基础结构。对于其他拓展功能，Flask则鼓励开发者使用外部库进行定制，其内核仅包含基础的网页模型。这种设计使得Flask具有了出色的可扩展性，这是其他Web框架难以比拟的。由于Flask的代码简洁明了，易于理解和维护，因此它能够很好地满足小型网站的开发需求。同时，Flask也展现出了强大的灵活性和可扩展性，它可以与各种数据库进行高效的交互，适应不同的项目需求。在真实生产环境中，Flask在小项目的开发上表现出了快速高效的特点，而在大项目中，由于其设计灵活，也能够提供足够的定制空间。因此，无论是小型网站还是大型项目，Flask都能够为开发者提供有力的支持。

2.5 数据集构建

2.5.1 中医古籍语料库构建

从公开资源中收集自秦汉以来的包括《黄帝内经》《金匮要略》《伤寒杂病论》《神农本草经》等经典中医古籍在内的608本中医古籍文本，结合机器校对和人工校对方法对古籍进行整理，并使用Python语言编写文本清洗、去重、合并脚本，去除文本中多余的HTML标签、广告等噪声信息，排除重复收集的文本，经过合并最终形成了一个83 032 043字的全面的、高质量的中医古籍简体文本语料库。中医古籍文本语料库涵盖的古籍种类及每种古籍涵盖的古籍数目和文字数量信息见表2.1。

表2.1 中医古籍文本语料库统计信息

古籍分类	古籍数目	古籍字数
本草	40	8 279 862
方药	79	20 217 503
妇幼	89	7 775 908
医论	151	23 169 590
伤科	52	4 599 361

经络	13	390 475
五官	24	938 615
针灸	26	1 332 989
医案	45	5 192 447
医经	76	10 397 047
四诊	13	738 246

2.5.2 马王堆简帛单字图像数据集构建

由于马王堆医书简帛文字年代久远，非专业的古文字研究人员，难以加以辨识。因此，很难为数据集添加标签。《长沙马王堆汉墓简帛集成》[50]由湖南省博物馆、复旦大学出土文献与古文字研究中心编纂，裘锡圭担任主编，提供了全面的简帛照片资料和扎实的释文注释基础。《马王堆汉墓简帛文字全编》[51]由《长沙马王堆汉墓简帛集成》一书的原班人马编纂，对所有简帛文字原始资料作了通盘的清理，秉持“有字必录”原则，收录了马王堆汉墓简帛中出现的3 339类文字。

扫描《马王堆汉墓简帛文字全编》，以PNG格式存储《马王堆汉墓简帛文字全编》的每一页照片。从《马王堆汉墓简帛文字全编》中人工截取100张马王堆简帛手写文字图片，用以训练YOLOv5目标检测模型，使用训练好的YOLOv5目标检测模型，自动检测并分割出《马王堆汉墓简帛文字全编》照片中的93 841张简帛文字图像。再基于《马王堆汉墓简帛文字全编》整理的字形，对分割出的所有单字图像进行人工分类整理，按照每个字单独存储为一个文件夹的形式进行存储，形成文本识别原始数据集。原始数据集共分为3 339类，但其各个类别的样本量极不均衡，样本数最大为3 086，最小仅仅为1，样本比例约为3 000：1，具体样本分布如图2.4所示。

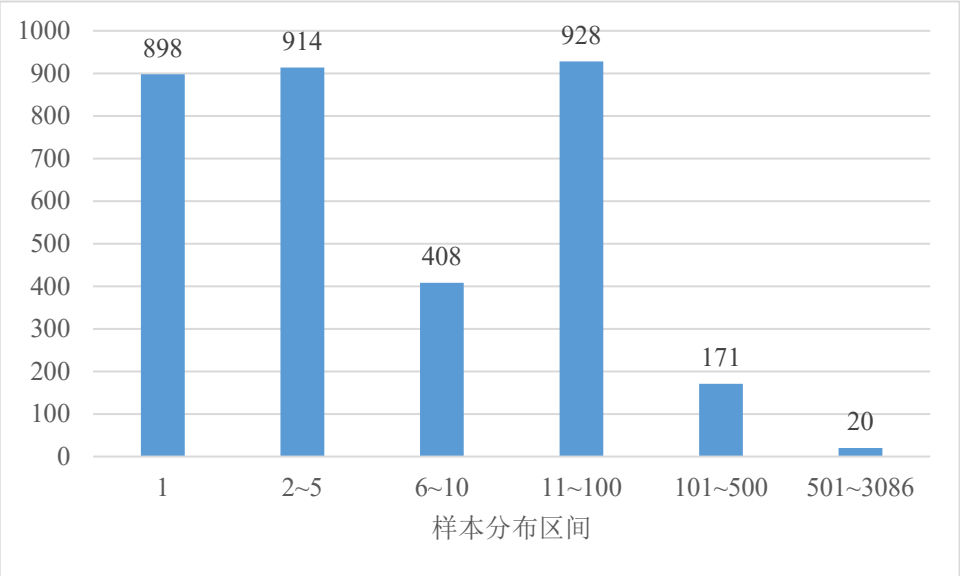

图2.4 原始数据集样本分布统计

为了解决数据不均衡导致模型泛化能力弱的问题，在原始数据集的基础上，使用BAGAN-GP网络学习马王堆简帛文字的字形特征，对出现次数较少的文字进行样本生成，生成结果如图2.5所示，其中第一列为原始图像，其余皆为BAGAN-GP网络生成的文字图像。在此基础上结合传统数据增强技术，包括随机旋转、随机缩放、加入噪声等方式来进一步平衡数据集。最终的单字数据集中每一类别的样本量约为3 000，总样本量为10 203 022。

图2.5 BAGAN-GP生成样本示例

2.6 本章小结

本章首先介绍进行中医古籍缺失文本预测和中医古籍简帛手写文字识别研究时所需的理论和技术。在中医古籍缺失文本预测方面介绍了语言模型技术的发展历程，重点介绍了BERT预训练语言模型，在古籍简帛手写文字识别方面介绍了生成对抗网络和CRNN算法及其评价指标。

然后介绍了中医古籍文本语料和简帛手写图像数据集的构建。着重介绍了如何使用生成对抗网络及传统数据增强方法扩充数据集以解决原始简帛单字图像数据集的数据不平衡问题。

第三章 中医古籍缺失文本预测研究

中医古籍是中医药学术与文化的重要载体，大量传世中医古籍对回溯学术源流、传承中医学术、深入研究中医药学理论、提高临床疗效有着不可替代的重要价值。因此，在中国特色社会主义新时代背景下，中医药学作为中国古代科学的瑰宝和打开中华文明宝库的钥匙，中医古籍的修复工作具有非凡的历史价值和文化价值。本章将基于语言模型技术，构建基于中医古籍的预训练BERT模型，实现对中医古籍中缺失的文本进行预测，从而辅助古籍修复专家的修复工作，提升其工作效率。

3.1 ACMBBERT 模型构建研究

3.1.1 BERT 模型

BERT模型是一种基于Transformer架构的自监督深层双向语言表示模型，它通过掩码语言模型迫使模型根据前后文全向信息进行预测，从而实现深层双向文本表示。如图3.1所示，在输入单条文本时，BERT模型通过查询字向量表将文本中的每个字转换为一维向量，将字向量、文本向量和位置向量的加和作为模型输入；模型输出则是输入各字对应的融合全文语义信息后的向量表示。

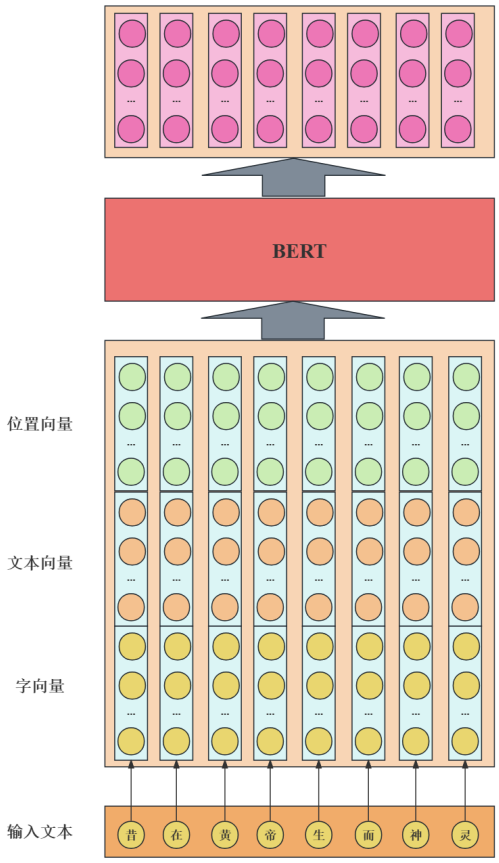

图3.1 BERT模型输入输出示意图

在MLM任务中，BERT模型在输入语句中随机掩盖掉15%的词，用其上下文来预测被掩盖的词。如图3.2所示，BERT对于在原句中被抹去的词汇，80%的情况下采用一个特殊符号[MASK]替换，10%情况下采用一个任意词替换，剩余10%情况下保持原词汇不变。从而强迫模型在编码当前时刻不能太依赖于当前的词，使模型更多地依赖于上下文信息去预测词汇。

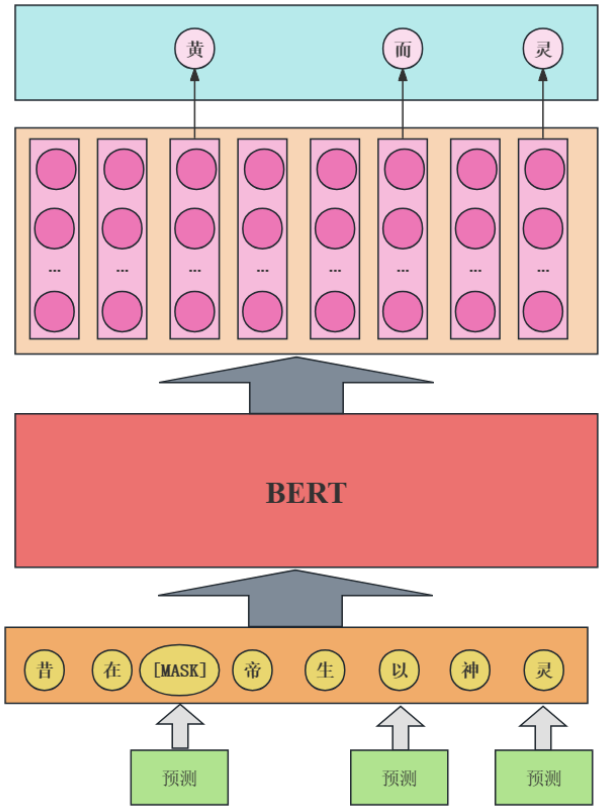

图3.2 MLM任务示意图

3.1.2 ACMBBERT 模型构建流程

图3.3展示了ACMBBERT模型的构建流程，分为三个部分，语料预处理，模型预训练，以及模型效果评测。实验在如下环境进行：CPU：Inter Core i7 11700K @ 3.60GHz、GPU：GeForce RTX 3090、内存：32GB，操作系统Windows10，64位，编程语言Python。在语料预处理阶段，收集整理了608本中医古籍语料并对其进行数据转化与清洗，再将其中的80%作为训练集，20%作为测试集。模型预训练阶段，在总结多次预实验结果后对训练参数进行调优，选取Huggingface提供的PyTorch版BERT-base-Chinese在训练集上使用MLM任务完成模型的预训练。在模型效果评测阶段，使用“hit@k”作为评价指标，判断预训练效果。

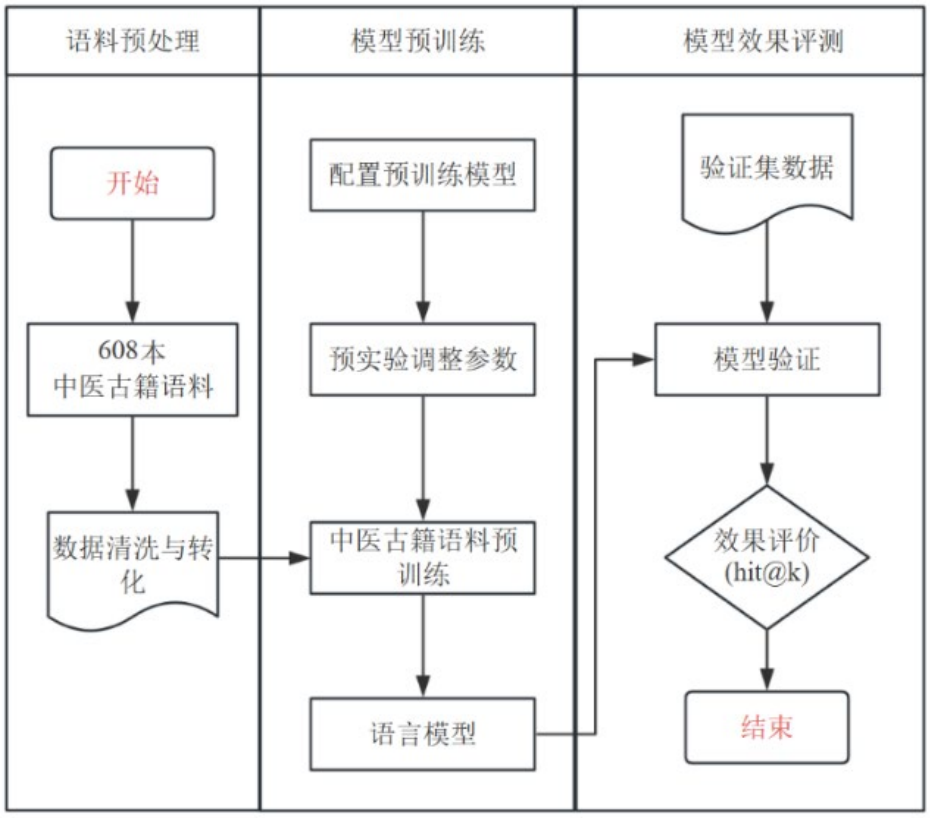

图3.3 ACMBBERT模型构建实验流程

在模型验证阶段，ACMBBERT模型在中医古籍语料库测试集上达到了60.79%hit@1和80.57%的hit@5以及hit@10，即ACMBBERT模型在中医古籍语料的MLM任务上有60.79%的几率可以直接获得正确的MASK内容，80.57%的几率在前10条预测中，出现正确的预测。

3.1.3 评价指标

本章使用“hit@k”作为缺失文本预测模型的评价指标，即在模型给出的前k个预测中，正确预测占所有句子的百分比。如果模型输出的正确预测排名越靠前，说明模型的性能越优秀。如表3.1中共有10条语料，模型的预测1即为真实值，那么hit@1即为100%。实际情况中，可能在第5个甚至第10个预测才会出现真实值，因此论文还会评估hit@5和hit@10，表示模型前5个和前10个预测中出现真实值的输出占测试语料总数的百分比。

3.2 中医古籍缺失文本预测研究

中医古籍缺失文本预测研究流程如图3.4所示。为了研究语料库的选择对中医古籍修复工作的影响，论文将中医古籍文本语料库划分为《黄帝内经》语料库、先秦两汉语料库，以及完整的中医古籍文本语料库。在划分完成的3个语料库的基础上分别对语料库进行分割，将其中的80%作为训练集、20%作为测试集。为了确定适用于中医古籍缺失文本预测的最优语言模型，本章基于中医古籍文本语料库及语言模型技术，在经过语料库划分的《黄

帝内经》语料库训练集、先秦两汉语料库训练集上分别训练N-gram模型，LSTM模型、BiLSTM模型和BERT模型，与3.1构建的ACMBBERT模型的实验结果进行对比，从而筛选出最优模型，并将筛选出的最优模型运用到文本修复场景中。

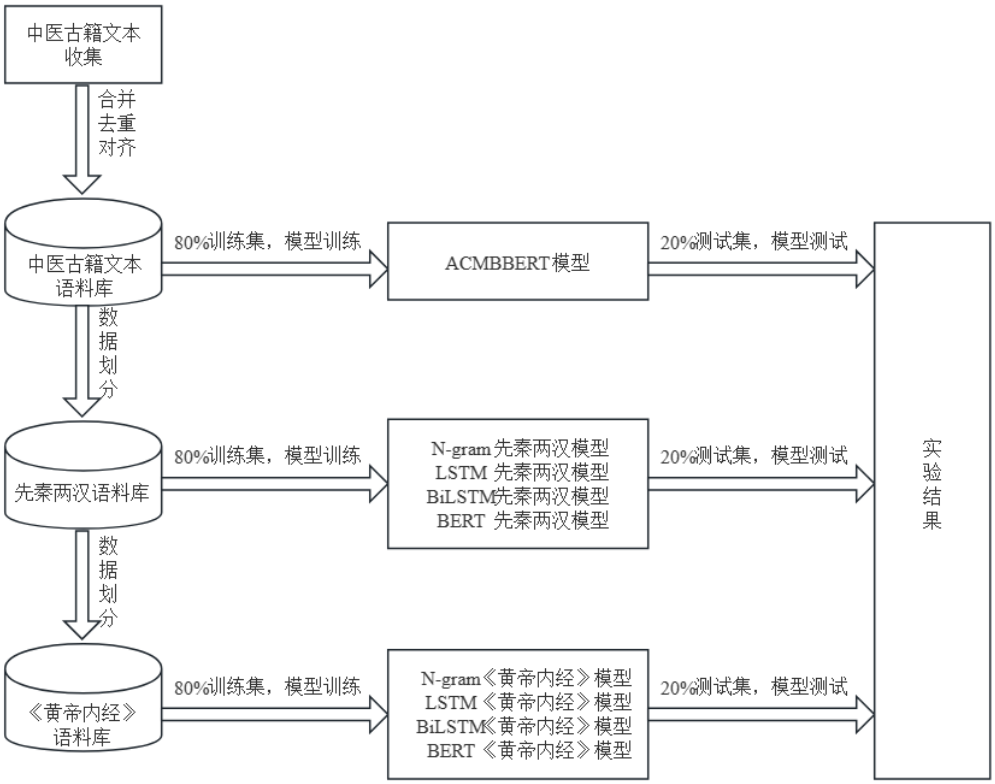

图3.4 中医古籍缺失文本预测研究流程

训练完成后，分别在对应的测试集上对训练出的各模型进行测试。测试语料的设置模拟文本修复场景，对古籍文本中的某条语句随机抹去其中的某一个字。图3.5为节选出的10条《黄帝内经》测试语料，一行一条语句，用“<mask>”标记每条语句中被随机抹去的字。《黄帝内经》测试集包含1 000条测试语料，先秦两汉测试集包含2 030条测试语料，中医古籍测试集包含180 623条测试语料，各模型将在对应的测试集上对每一条测试语料输出尽可能多的预测。

```
征 其 脉 不 夺 其 <mask> 夺 者 ， 此 久 病 也 ；
帝 曰 ： 阴 与 阳 并 ， 血 气 以 并 ， 病 形 以 成 ， 刺 之 <mask> 何 ？
足 少 阴 令 人 <mask> 痛， 痛 引 脊 内 廉 。
取 其 经 太 阴 、 阳 明 、 少 阴 血 <mask> 。
颇 痛 ， 刺 手 阳 明 与 颇 之 盛 脉 出 <mask> 。
刺 缺 盆 中 内 陷 气 <mask> ， 令 人 喘 咳 逆 。
热 之 而 寒 者 ， <mask> 之 阳 ；
若 有 七 诊 <mask> 病 ， 其 脉 候 亦 败 者 死 矣。
岐 伯 曰 ： 无 者 生 <mask> ， 有 者 甚 之。
黄 帝 曰 ： 取 之 有 <mask> 乎 ？
```

图3.5 《黄帝内经》测试语料节选

为了评估模型在真实中医古籍文本修复场景的应用效果，在没有专业的中医文本修复

专家的帮助下直接去修复一本残损的古籍，模型的预测是无法进行参照对比的，而《黄帝内经》的发行版本被广泛接受认可，因此挑选了《黄帝内经》作为修复对象。在整个中医古籍文本语料库中剔除《黄帝内经》原文及和《黄帝内经》直接相关的古籍，将其命名为“中医古籍-S”语料库，在“中医古籍-S”语料库上重新训练ACMBBERT模型，将其命名为“ACMBBERT-S”模型，以《黄帝内经》的文本作为测试语料构建测试集进行测试。

为了测试模型的泛化能力，在整个中医古籍文本语料库中剔除马王堆医书相关语料，将其命名为“中医古籍-MWD”语料库，在“中医古籍-MWD”语料库上重新训练ACMBBERT模型，将其命名为“ACMBBERT-MWD”模型，以马王堆医书的已知文本作为测试语料构建测试集，测试模型的泛化能力。

3.3 结果分析

表3.1为模型对图3.5中节选的10条测试语料的输出结果，如对第一条语料“征其脉不夺其<mask>夺者，此久病也”，模型输出“色，0.999891639”、“脉，0.0000349873371305875”和“目，0.00002899141691159 45”，即模型预测此处的“<mask>”可能为“色”“脉”“目”，同时“色”是真实值的概率高达99.98%。实际情形中，模型会给出至少5个预测，以及预测为真实值的概率。

表3.1 模型预测示例

预测1	正确概率	预测2	正确概率	预测3	正确概率
色	0.999891639	脉	0.0000349873371305875	目	0.0000289914169115945
奈	0.999989629	如	5.30911893292795E-06	谓	8.90942146725137E-07
腰	0.999875069	少	0.0000221588470594724	心	0.0000164736065926263
者	0.999988079	也	5.73332863496034E-06	络	3.42004818776331E-06
血	0.999992609	者	3.53919358531129E-06	止	9.41970824896998E-07
泄	0.99992466	逆	0.0000336658849846571	竭	0.0000069860293478996
取	0.999988317	补	5.81890344619751E-06	泻	1.50885739458317E-06
之	0.999897003	病	0.0000187704736163141	者	0.0000169722879945766
之	0.999999881	气	7.79439943698889E-08	神	2.05727399560373E-08
数	0.996884644	道	0.00293516251258552	名	0.0000466143937956076

表3.2和表3.3显示了各模型在对应训练集和测试集上的表现。对表3.2和表3.3的结果进行各模型间的横向对比可以发现，BERT模型效果最优，BiLSTM模型优于LSTM模型，LSTM模型明显优于N-gram模型；对表3.2和表3.3的结果进行各语料库间的纵向对比，可以发现当语料库规模越大，模型的预测效果越优，说明当中医古籍修复专家进行修复工作时，其研究的中医古籍范围越广，程度越深，也就越能推测出中医古籍中缺失的内容，古籍修复工作效果也就越好。

表3.2 各模型训练集预测效果

语料库	模型	评价指标		
		hit@1/%	hit@5/%	hit@10/%
《黄帝内经》训练集	N-gram《黄帝内经》模型	50.45	66.42	71.34
	LSTM《黄帝内经》模型	79.00	94.00	94.00
	BiLSTM《黄帝内经》模型	68.03	81.06	84.36
	BERT《黄帝内经》模型	99.40	99.90	99.90
先秦两汉训练集	N-gram 先秦两汉模型	51.31	67.55	72.25
	LSTM 先秦两汉模型	75.11	93.01	95.99
	BiLSTM 先秦两汉模型	75.34	93.53	96.55
	BERT 先秦两汉模型	97.50	99.40	99.40
中医古籍训练集	ACMBBERT 模型	99.50	99.90	99.90

表3.3 各模型测试集预测效果

语料库	模型	评价指标		
		hit@1/%	hit@5/%	hit@10/%
《黄帝内经》测试集	N-gram《黄帝内经》模型	22.52	41.44	45.94
	LSTM《黄帝内经》模型	23.63	36.36	47.27
	BiLSTM《黄帝内经》模型	28.18	35.45	42.72
	BERT《黄帝内经》模型	39.00	56.80	56.80
先秦两汉测试集	N-gram 先秦两汉模型	32.03	50.89	54.49
	LSTM 先秦两汉模型	37.12	54.19	58.38
	BiLSTM 先秦两汉模型	38.32	54.79	60.47
	BERT 先秦两汉模型	51.70	70.10	70.10
中医古籍测试集	ACMBBERT 模型	60.79	80.57	80.57

在剔除《黄帝内经》原文及《黄帝内经》直接相关古籍的“中医古籍-S” 语料库上重新训练的ACMBBERT-S模型，最终可以达到63.36%的hit@1，82.57%的hit@5和hit@10，即随机在《黄帝内经》中剔除一个字，模型有63.36%的概率在其第一个预测中给出真实预测，有82.57%的概率在其给出中的前5个或前10个预测中出现真实预测。实验结果显示，ACMBBERT-S模型的模拟修复效果与ACMBBERT模型最终的测试效果相符合，说明了ACMBBERT-S模型的实用性。

同时，在一些专家已经给出古籍中真实缺字预测的情况下，ACMBBERT-S模型也能给出和专家相近的预测。如1973年长沙马王堆出土的《五十二病方》中伤痉第六方为“冶黄芩、甘草相半，即以鼬膏财足以煎之。煎之沸，即以布捉之，抒其汁，□傅□”，其中“□”代表缺字，模型对第一个“□”的预测为“以”“勿”“即”，第二个“□”的预测为“之”“讫”“头”，而周一谋先生在《马王堆医书考注》[52]中对第一个“□”也给出了“以”的预测，严健民先生在《五十二病方注补译》[53]中对第二个“□”同样给出了“之”的预测；同样于马王堆出土的《足臂十一脉灸经》中臂少阳脉“出中指，循臂上骨下廉，凑耳。其病：产聋，□痛。诸病此物者，皆灸臂少阳之脉”，模型对此处“□”给出的预测为“胁”“耳”“头”“颊”，周一谋先生在《马王堆医书考注》中认为此处的“□”应为“颊”。

而在剔除了马王堆医书相关语料的“中医古籍-MWD”语料库上训练的ACMBBERT-

MWD模型，最终可以达到37.58%的hit@1，51.42%的hit@5和57.44%的hit@10，即在已知的马王堆医书语料库上随机剔除一个字，ACMBBERT-MWD模型有37.58%的概率在其第一个预测中给出正确预测，有51.42%的概率在前五个预测中给出正确预测，57.44%概率在前十个预测中出现真实预测。与ACMBBERT-S模型相比，ACMBBERT-MWD模型在预测马王堆医书中缺失文本的准确率虽然不及ACMBBERT-S模型预测《黄帝内经》的缺失文本，但是也说明了模型具有了一定的泛化能力。

实验结果证明了本文构建的ACMBBERT模型在中医古籍缺失文本修复任务上的实用性，ACMBBERT-S模型和ACMBBERT-MWD模型的验证实验表明，ACMBBERT模型同时具备一定的泛化能力和推广性，可以为后续的新出土的残损中医古籍的修复提供参考，辅助中医古籍修复专家进行中医古籍修复工作。

3.4 本章小结

本章主要介绍了ACMBBERT模型的构建流程、评价指标和中医古籍缺失文本预测研究的实验流程。在自建的中医古籍语料库上训练了各语言模型，验证了ACMBBERT模型在缺失文本预测任务上的优势。将ACMBBERT模型运用到中医古籍文本修复场景中，分别训练了ACMBBERT-S模型和ACMBBERT-MWD模型，其中ACMBBERT-S模型评hit@1为63.36%， hit@5和hit@10均为82.57%，ACMBBERT-MWD模型hit@1为37.58%，hit@5为51.42%，hit@10为57.44%，达到了训练一个中医古籍文本修复模型的预期目标，可为中医古籍修复专家提供参考和支持。

第四章 中医古籍简帛手写文字识别研究

随着国家高度重视中医药事业的发展，关于中医药的学术研究，特别是关于中医药的继承和创新研究已成为国内外关注的一个焦点问题。长沙马王堆出土的古医书既是一个中医药宝库，也是中国文化的重要代表之一。马王堆医书对于中医理论之大成者《黄帝内经》有奠基之功，可以称之为后世中医理论及养生理念的源头活水。虽然扫描拍摄和人工输入的方式使我国古籍数字化工作取得了长足的进步，《长沙马王堆汉墓简帛集成》整理了马王堆出土简帛的高清照片，《马王堆汉墓简帛文字全编》归纳汇总了简帛中出现的所有文字，但是对于马王堆简帛的研究人员来说，阅读简帛原始内容时，由于简帛残损严重，年代久远，与现今字体差异过大等原因，仍然难以直接辨识简帛中的文字。

因此本章将基于深度学习技术，从马王堆简帛单字数据集的构建、马王堆简帛图像单字识别、结合语言模型纠错的马王堆医书简帛图像长文本序列识别等方面开展研究，实现马王堆医书简帛图像手写文本的自动识别，从而响应国家号召，助力古籍数字化，为中国文化的传承与延续贡献力量。

4.1 马王堆简帛手写单字识别研究

马王堆简帛的手写单字识别问题本质上是一个图像分类的问题，关键点在于构建一个3 339类的分类模型，提升模型的泛化能力，解决其在常见字的过拟合和非常见字的欠拟合问题。

为了研究使用生成对抗网络和传统数据增强方法平衡数据集的方式对文字识别模型准确率提升的效果，论文基于图像分类领域常见数据集ImageNet[54]上各个网络模型的表现，对于多个基础图像分类网络进行研究，最终选择ResNet[55]作为基础网络模型，在原始数据集和平衡数据集上进行马王堆汉墓简帛文字的识别对比实验。

此外，为了研究预训练权重对模型泛化能力提升的影响，论文对ResNet、DenseNet[56]、MobileNet[57]、ShuffleNet[58]和EfficientNet[59]这五个模型分别进行从头训练以及迁移学习训练，比较两种不同方式训练的分类模型的准确率和收敛速度。

4.1.1 数据集

马王堆简帛手写文字初始数据集包括3 339类，然而初始数据集样本数据极其不平衡，其中最多的“之”字有3 086个样本，而最少的字仅有一个样本。数据不平衡的问题将导致模型对小样本类别文字识别的欠拟合，对大样本类别文字识别的过拟合，从而影响整体的预测能力和性能。因此论文在原始数据集基础上训练了一个基于BAGAN-GP图像生成模型，对原始数据集进行扩充。通过生成对抗网络的样本生成，数据集的不平衡性在一定程度上得到了缓解。每一类补充500个样本，样本数量上的差距仍然存在。因此论文通过随机旋转、随机缩放、加入噪声等方式来进一步平衡数据集。最终，数据集中每一类别的图片数量约

为3 000张，总共样本图片数为10 203 022张。

4.1.2 评价指标

马王堆简帛文字识别是一个多分类问题，多分类问题中常使用微平均Precision、微平均召回率和微平均F1-score作为评价指标，而微平均Precision、微平均召回率和微平均F1-score三者在数值上与准确率Accuracy相等，因此本文使用Accuracy作为评价指标，具体计算公式如下。

设n为类别数，TP、FN、FP、TN表示True Positive、False Negative、False Positive、True Negative样本的数量。

准确率计算公式如式4-1：

$$\mathrm{Accuracy}=\frac{\sum_{i=1}^{n}\mathrm{TP_i}}{\sum_{i=1}^{n}(\mathrm{TP_i}+\mathrm{FN_i})} \tag{4-1}$$

微平均精确率计算公式如式4-2：

$$\mathrm{Precision_{micro}}=\frac{\sum_{i=1}^{n}\mathrm{TP_i}}{\sum_{i=1}^{n}(\mathrm{TP_i}+\mathrm{FN_i})} \tag{4-2}$$

微平均召回率的计算公式如式4-3：

$$\mathrm{Recall_{micro}}=\frac{\sum_{i=1}^{n}\mathrm{TP_i}}{\sum_{i=1}^{n}(\mathrm{TP_i}+\mathrm{FN_i})} \tag{4-3}$$

微平均F1-score的计算公式如式4-4：

$$\mathrm{F1-score_{micro}}=\frac{2\mathrm{Precision_{micro}}\times\mathrm{Recall_{micro}}}{\mathrm{Precision_{micro}}+\mathrm{Recall_{micro}}}=\mathrm{Accuracy} \tag{4-4}$$

4.1.3 实验设计与结果分析

马王堆古简帛文字的识别模型的训练环境均为Ubuntu操作系统，CPU为Intel(R)Xeon(R) Platinum 8255C，内存40GB，GPU为NVIDIA V100，显存大小为32GB。

由于文字识别是一个多分类问题，因此选择交叉熵作为损失函数。优化器选择Adam，其能够按照目前的真实情况给出最合适的学习率，从而加快模型收敛，并且能够很好的处理大规模数据集。具体超参数见表4.1。

表4.1 ResNet超参数设置

	原始数据集	平衡数据集
学习率	0.001	0.001
优化器	Adam	Adam
迭代次数	20	20
批大小	64	64
损失函数	交叉熵误差	交叉熵误差

迁移学习对比实验用于对比从头训练与迁移学习之间的效果，具体超参数设置见表4.2。

表4.2 迁移学习实验超参数设置

训练参数 / 训练方式	学习率	优化器	迭代次数	批大小	损失函数
从头训练	0.001	Adam	20	64	交叉熵误差
迁移学习	0.001	Adam	20	64	交叉熵误差

在经过BAGAN-GP和数据增强补充样本后的平衡数据集上训练的ResNet分类模型整体性能更高。如表4.3所示，基于平衡数据集训练的ResNet模型在测试集和验证集上的表现都要优于基于原始数据集训练的ResNet模型。此外，原始数据集中由于样本量小，必定会识别错误的字，例如“刖”字，在平衡数据集上可以被准确识别。因此，数据集的不平衡性确实会影响分类模型的整体准确率，而本文的BAGAN-GP和传统数据增强方法结合优化数据集的方式可以用于解决马王堆古简帛数据集的不平衡问题，生成的样本对模型的训练可以起到积极作用。

表4.3 生成图片实验准确率对比

数据集来源 / 模型	**测试数据集**	**检验数据集**
原始数据集训练的 ResNet	89.408%	72.770%
平衡数据集训练的 ResNet	98.589%	98.328%

如图4.1、图4.2和表4.4所示，图例中“_pre”表示以迁移学习方式训练的模型，“_unpre”表示以从头训练方式训练的模型，虽然迁移学习和从头训练两种方式得到的分类模型在测试集的准确率差距不大，但迁移学习方式训练的模型在训练过程中在测试集上的准确率和损失值都能更快的提升并且达到收敛，且整体的准确率都要略高于从头训练的模型。根据模型大小以及测试集准确率，本文选择EfficientNet V2 small作为最终的马王堆汉墓简帛文字的单字识别分类模型。

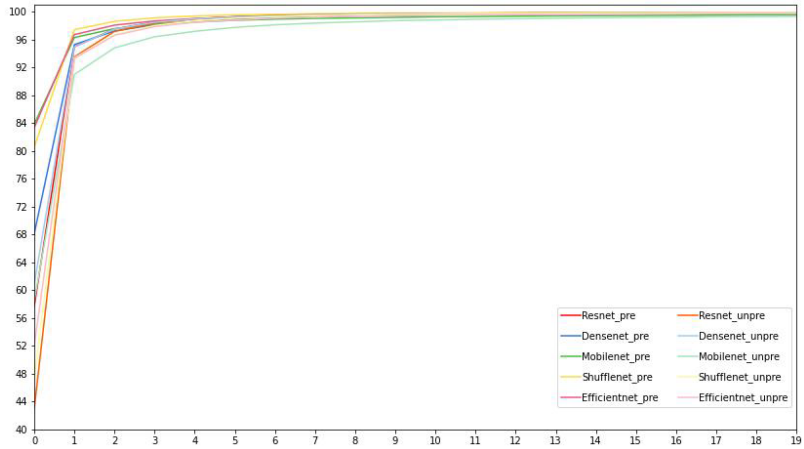

图4.1 各模型测试集准确率对比

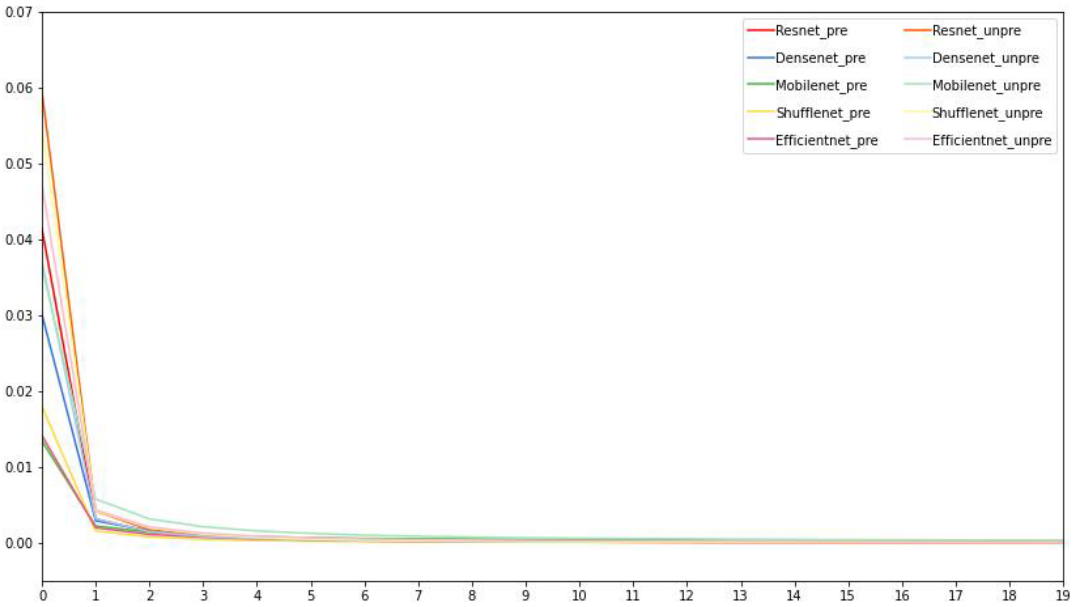

图4.2 各模型测试集loss对比

表4.4 从头训练和迁移学习准确率一览表

训练方式 / 分类模型	从头训练		迁移学习	
	测试集准确率	模型大小	测试集准确率	模型大小
ResNet152	98.589%	248.8M	98.673%	248.8M
DenseNet201	98.743%	94.8M	98.598%	94.8M
MobileNet V3 small	98.005%	19.0M	98.405%	19.0M
ShuffleNet V2 x2.0	98.677%	46.7M	98.871%	46.7M
EfficientNet V2 small	**98.874%**	**94.2M**	**99.048%**	**94.1M**

马王堆简帛手写文字的单字识别实验结果表明，BAGAN-GP 生成模型与传统数据增强方法结合的方式能够很好地适用于马王堆简帛手写文字数据集的扩充。扩充后的平衡数据集在不同的图像分类网络上都可以取得很高的识别准确率。此外，结合迁移学习方式，导入预训练权重，模型的训练可以更快地收敛，准确率也可以得到一些提升。

4.2 简帛手写文本序列识别研究

马王堆简帛手写文字的单字识别研究可以较为准确地对马王堆出土简帛中所有出现的文字进行分类，同时也解决了马王堆手写文字序列识别研究的数据来源问题，但是从实际应用场景中看，仅仅对单字进行识别，并不能高效地助力研究人员识别阅读古籍。因此简帛手写文本序列识别研究致力于解决长序列文本的识别问题，提高简帛文本识别效率和

准确率，并提出一种高效的，易复用的古籍文本识别流程，如图 4.3 所示。首先使用 YOLOv5[60]目标检测模型分割出《全编》中的所有单字图片，然后进行人工分类整理，组成马王堆简帛不平衡手写单字数据集，接着使用 BAGAN-GP 网络和传统数据增强方法对不平衡数据集进行扩充，形成马王堆简帛平衡手写单字数据集。在平衡单字手写数据集基础上，从每一类中随机抽取一张图片，拼接成随机长度的文本列，构建长文本图像数据集，作为马王堆简帛手写长文本识别模型的训练数据。如此，可以生成大量长度不一的训练数据，无需从马王堆简帛原始照片中截取文本列，再不断地人工查阅《全编》添加标签，极大提升数据集构建的效率。再使用从《长沙马王堆简帛集成》中随机分割出的文本列图像进行测试，测试基于生成数据集训练的马王堆序列识别模型在真实场景下的性能。接着，将基于生成数据集训练的马王堆简帛手写长文本识别模型的输出输入到基于 ACMBBERT 模型和马王堆医书文本语料微调的马王堆医书文本纠错模型中，对模型输出中可能存在的错误进行纠正以提升马王堆简帛手写长文本识别模型识别结果的准确率。

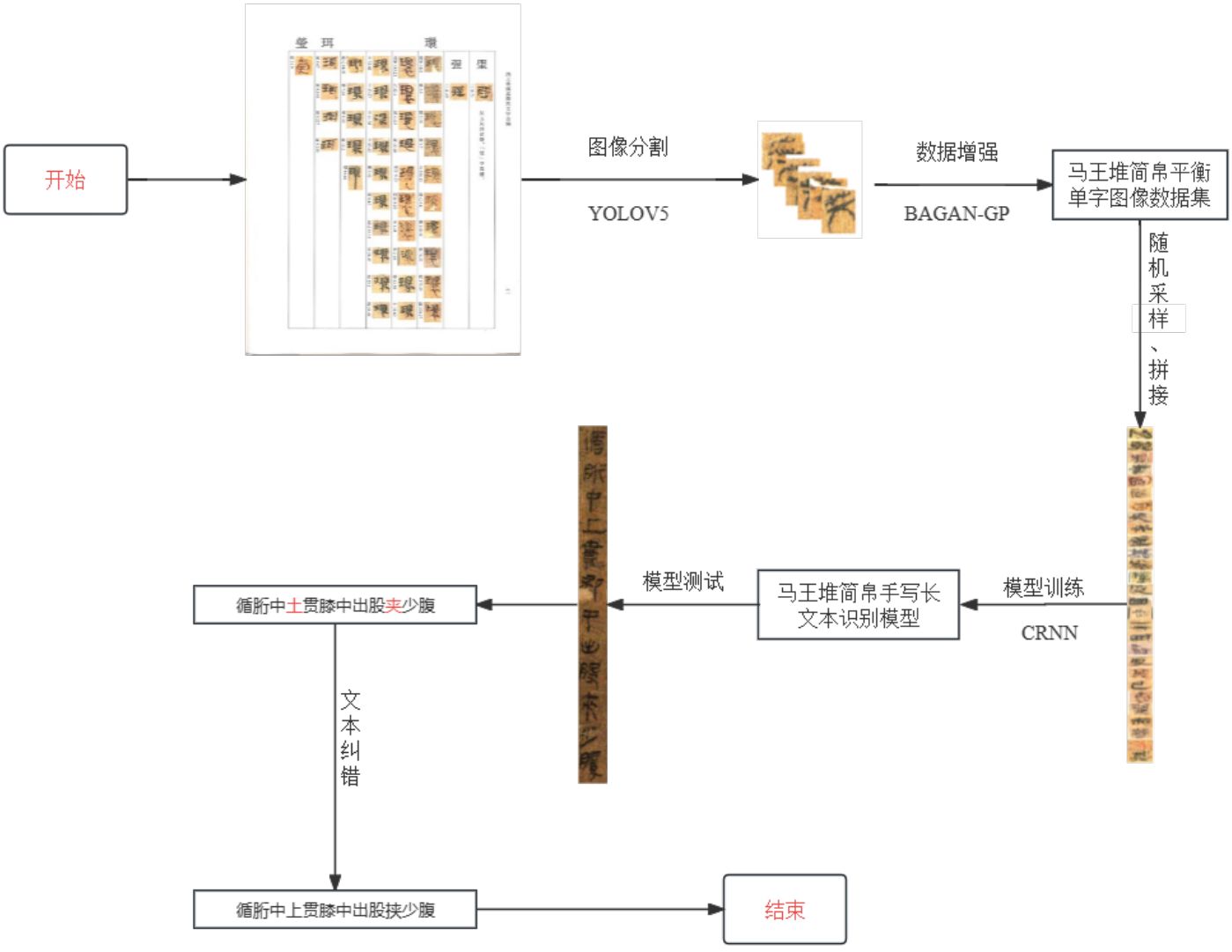

图4.3 简帛手写文本序列识别研究流程

4.2.1 评价指标

深度学习的常用评价指标有准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1值（F1 score)。马王堆医书简帛文本序列识别模型关注模型对帛书图像文本的识别能力，因此使用准确率作为评价指标。文本纠错模型分为检测网络和纠错网络，模型将预测结果与真实值做判断，本质上是一个二分类问题。因此论文在字符级别同时关注检测网络和纠错网络的性能，使用召回率、精确率、F1值作为评价指标，统计模型能检测出错字和正确纠正错字的能力，在句子级别关注模型对整个句子进行文本纠错的能力，在字符级别的基础上增加准确率作为评价指标，统计模型能检测出句子存在错字和完全正确修改句子的能力。

TP、FN、FP、TN表示True Positive、False Negative、False Positive、True Negative样本的数量。

准确率计算公式如式4-5：

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{4-5}$$

精确率计算公式如式4-6：

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{4-6}$$

召回率的计算公式如式4-7：

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{4-7}$$

F1-score的计算公式如式4-8：

$$\text{F1-score}=\frac{\text{2*Precision*Recall}}{\text{Precision+Recall}} \tag{4-8}$$

4.2.2 实验设计与结果分析

马王堆简帛手写文字识别属于OCR（光学字符识别，Optical Character Recognition）的范畴，与通用领域OCR的区别在于马王堆简帛文本内容的特殊性。马王堆出土简帛的残损状况较为严重，书体为篆隶，多为异体字，同时还是手写体，缺乏相关的数据集。但是，马王堆医书简帛文本多为规则形状的文本列，而OCR领域中的CRNN算法在规则文本的识别上广泛应用。因此研究首先基于构建的生成数据集和CRNN算法，训练马王堆简帛长文本识别模型。

虽然马王堆简帛长文本识别模型在生成数据集和由真实马王堆简帛图像构成的测试集上取得了不错的表现，在生成数据集上的准确率达到了86.73%，在测试集上的准确率也

达到了63%，但是准确率方面仍有提升空间。基于语言模型纠错的马王堆医书简帛长文本序列识别模型在马王堆简帛长文本识别模型的基础上，以马王堆简帛长文本识别模型的输出作为输入，自动检测输入文本的错误文字，例如输入“足臂十二脉灸经”，语言模型可以检测出“二”是一个错误，并基于此处错误的上下文，预测此处的“二”为“一”，最终输出正确的预测“足臂十一脉灸经”，实现对马王堆简帛长文本识别模型输出优化的效果，从而提升文本识别模型识别的正确率。

在大部分情况下，马王堆简帛长文本识别模型对测试集的预测是准确的，即使错误发生，一条10个字的语句，也只会发生个别字的错误预测。因此，错误预测发生时，错字的上下文信息是比较完整的，即可以根据错字上下文信息，对错字进行纠正。在缺失文本预测研究中，论文已基于构建的83 032 043字的中医古籍数据集，构建了ACMBBERT模型，该模型可以较为准确地根据缺失文字的上下文信息预测出一条中医古籍文本中缺失的文字。因此，论文基于PyCorrector下的MacBERT4CSC（MacBertMaskedLM for Chinese Spelling Correction）[61]架构使用马王堆医书文本对ACMBBERT模型进行微调，构建马王堆医书的文本纠错模型。将马王堆简帛长文本识别模型的输出作为马王堆医书文本纠错模型的输入，测试真实情形下马王堆医书纠错模型的性能。其中，MacBERT4CSC是在MacBERT（MLM as correction BERT）[62]基础上改变网络结构的中文文本纠错模型，加入了错误检测网络，利用检测层和纠错层的loss加权得到最终的loss，较好地适配了中文纠错任务，在通用领域上有着良好的纠错效果。

实验在如下环境进行：CPU：Inter Core i7 11700K @ 3.60GHz、GPU：GeForce RTX 3090、内存：32GB，操作系统windows10，64位，编程语言Python。实验分为三阶段，在预处理阶段，将数据尺寸大小统一调整为32×512，并做灰度变换，在网络训练阶段，选择MobileNetV3作为骨干网络，在后处理阶段使用CTC的解码方式，优化器为Adam，学习率为0.0 005。

表4.5 文本纠错模型性能

数据集	模型	字符级			句子级			
		精度/%	召回率/%	F1 值/%	准确率/%	精度/%	召回率/%	F1 值/%
训练集	检错模型	99.599	99.971	99.785	98.412	98.412	99.214	99.2
	纠错模型	99.598	99.569	99.583				
验证集	检错模型	99.735	98.683	99.867	99.134	99.743	99.133	99.57
	纠错模型	99.863	99.874	99.868				
测试集	检错模型	99.584	99.967	99.775	98.236	99.817	92.236	99.11
	纠错模型	99.518	99.551	99.534				

表4.5为马王堆医书文本纠错模型在生成马王堆文本纠错数据集上的表现，可见基于ACMBBERT模型和马王堆医书文本语料库微调的马王堆医书文本纠错模型在字符级和句子级都有着很好的性能。表明模型不仅可以准确地检测出句子中的错字并给出正确的预测，而且可以准确地判断句子是否存在错字并完全纠正句子。

表4.6为两种文本识别模型在训练集和测试集上的表现。由生成数据集训练的马王堆简帛长文本识别模型对生成的马王堆简帛图像识别准确率达到了86.73%，而由于马王堆医书简帛的真实照片和生成数据在颜色、纹理上还是存在一些差异，所以模型在测试集上的表现并不理想。100张测试图片中，仅仅有63张做到了完全正确的预测。

表4.6 文本识别模型性能

模型	数据集	准确率
马王堆简帛长文本识别模型	生成数据集	86.73%
	测试集	63.00%
结合语言模型纠错的马王堆医书长文本识别模型	测试集	91.00%

对马王堆简帛长文本识别模型错误的预测情况进行分析，发现在错误预测中，虽然某些字的预测发生了错误，但并不影响整条语句语义的理解。因此，正如表4.6中实验结果所示，结合马王堆医书文本纠错模型对这些错误进行纠正，在一定程度上提升了文本识别模型的识别准确率。同时，由于生成数据集是随机生成的，其真实标签之间不存在语义关系，因此不考虑文本纠错模型在生成数据集上对识别模型准确率的提升，而在100张由马王堆医书简帛真实图片构成的测试集上，马王堆医书文本纠错模型较大程度地提升了马王堆简帛长文本识别模型在马王堆医书相关简帛真实图像上的准确率。

此外，为了验证模型的泛化能力，论文还从成都老官山汉墓出土的《天回医简》[63]中截取了50幅图像，用以验证基于马王堆简帛平衡单字数据集字生成的长文本数据集训练的马王堆简帛长文本识别模型是否具备识别同时期其他出土古籍简帛文字的能力。50幅图像包含25幅单字图像和25幅长文本图像，其中长文本图像最长的含8个字，最短的含2个字，共计134个字。对于25幅单字图像，马王堆简帛长文本识别模型能够准确地识别出其中19幅图像，而对于另外25幅长文本图像，虽然模型能够完全准确地识别出的仅有4幅，但是模型对25幅图的预测中最多仅识别错3个字，且仅有一个字识别错误的有12幅。同时模型共计识别出了134个字中的81个字，在字符上达到了60.44%的准确率。如图4.4所示，即使是错误识别的情况，模型给出的预测也并没有脱离字形的限制。例如“首领”的“领”字被模型识别为了“履”字，“阳明各五”中的“明各”被模型认为是一个整体，识别为了“罌”，“东北风之风”中第一个风被识别成了“骨”。进一步对比“履”、“罌”和“骨”和“令”、“明各”、“风”的书写方式，可以发现，这些字在写法上是有相似之处的。证明了马王堆简帛长文本识别模型学习

到了秦汉时期篆隶体的手写特征，具备一定的泛化能力。

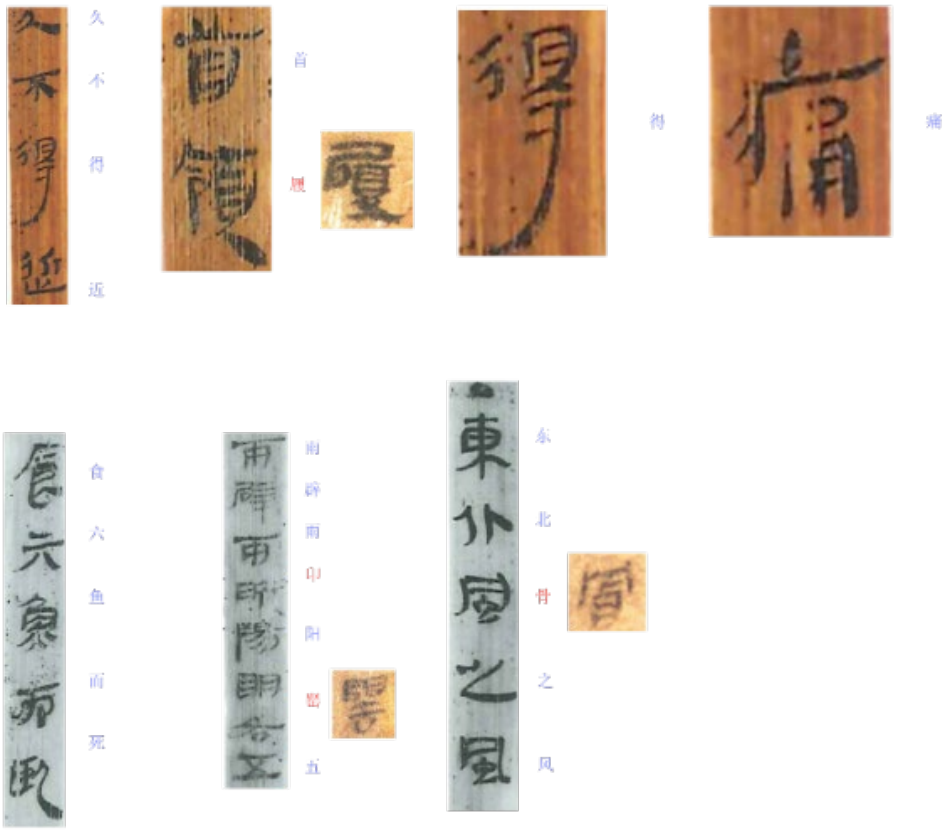

图4.4 《天回医简》部分简帛图像识别情况

4.3 本章小结

本章选取马王堆简帛文字为研究对象，利用深度学习技术，构建马王堆简帛文字识别网络，将BAGAN-GP生成对抗网络与传统数据增强方法结合，解决了古籍文字识别时经常出现的数据不平衡问题，并通过ResNet构建图像分类模型验证该数据增强方式的可行性，再结合迁移学习导入预训练权重提升模型的泛化能力和收敛速度。接着基于构建的平衡单字数据集和CRNN算法，随机生成了大量的文本序列图像数据，训练马王堆简帛手写长文本识别模型。解决了基于深度学习方法对古籍进行数字化时缺乏相关数据集或数据标注困难的问题，提升了数据集构建的效率。再基于马王堆医书文本对ACMBBERT模型进行微调，构建了马王堆医书文本纠错模型。并成功将马王堆医书文本纠错模型与马王堆手写长文本识别模型进行结合，从图像和文本两个角度，提升了马王堆医书简帛图像长文本识别的准确率。

第五章 中医古籍缺失文本预测和手写文字识别系统的设计和实现

本章基于前几章的研究内容和方法，使用Flask框架和HTML、CSS和JavaScript技术设计并实现了网页版的中医古籍缺失文本预测和手写文字识别系统，系统由首页及三个主要功能模块组成，分别为缺失文本预测模块、马王堆简帛手写单字图像识别模块和马王堆医书长文本序列识别模块。用户可以在网页中便捷地使用本文构建的中医古籍缺失文本预测模型和简帛手写文字识别模型。

5.1 中医古籍缺失文本预测和手写文字识别系统的整体框架

5.1.1 系统需求分析

系统设计目的是方便古籍研究人员使用本文构建的中医古籍缺失文本修复模型，马王堆简帛手写单字识别模型以及马王堆医书长文本序列识别模型，能够自行调用模型进行缺失文本的预测或马王堆简帛图像手写文字的识别并将检测结果返回给前端页面。综上，本系统的需求可归纳为以下几点：

（1）支持用户输入单条或多条文本语料。

（2）支持用户上传待识别的图像。

（3）能够将算法模型的预测结果返回前端，并在页面上显示。

5.1.2 系统总体设计

系统整体框架如图5.1所示，用户于前端网页上传待测古籍文本或待识别图像，服务器接收用户上传的数据，并调用缺失文本预测模型或文本识别模型对用户上传的文本、图像进行预测，并将预测结果通过服务器发回前端于网页显示。

图5.1 系统整体框图

中医古籍智能修复系统包含一个首页以及三个功能模块页，缺失文本预测模块、马王堆简帛手写单字识别模块和马王堆医书长文本序列识别模块。首页支持用户自由地选择所需进行的任务，缺失文本预测模块支持用户输入以“<mask>”标记缺失文本的语句，输出缺失文本预测模型对“<mask>”的多个预测及概率。马王堆简帛手写单字识别模块支持用户输入一张简帛单字图片，输出马王堆简帛单字识别模型的识别结果。马王堆医书长文本序列识别模块支持用户上传文本列图片，并返回识别结果。

5.2 中医古籍缺失文本预测和手写文字识别系统实现

保证服务器开启的情况下，通过前端网页提交用户需要预测的中医古籍缺失文本或需要识别的马王堆简帛手写文字图片，网页接收用户提交的语料和上传的图片，后端程序根据具体任务调用对应的算法模型，完成对缺失文本的预测或马王堆简帛手写文字的识别，最后将识别结果传回到前端网页中。后端程序预测过程如图 5.2 所示。

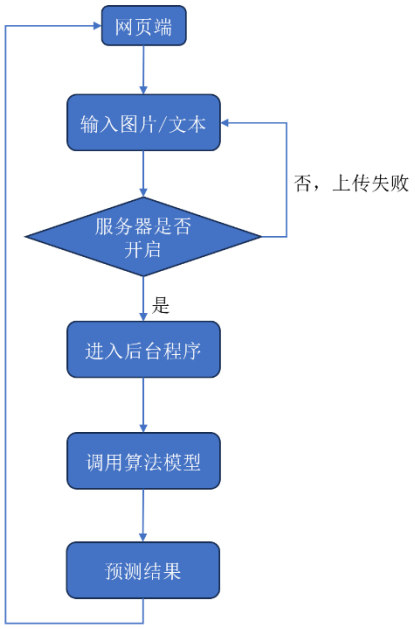

图5.2 后端程序运行过程图

5.2.1 首页功能实现

首页实现效果如图5.3所示，以超链接的形式提供缺失文本预测模块，马王堆简帛手写单字识别模块和马王堆医书长文本序列识别模块的入口，以便用户快捷地使用论文实现的三种算法。

图5.3 首页

使用Flask框架的render_template方法渲染超链接标签指向的功能模块页面，以下为首页跳转各功能模块页的后端实现代码：

```
@app.route('/中医古籍缺失文本修复系统.html')
def index1():
    return render_template('中医古籍缺失文本修复系统.html')
@app.route('/马王堆简帛手写单字图像识别系统.html')
def index2():
    img_stream = return_img_stream("./static/马王堆简帛手写单字图像识别系统_files/default.png")
    return render_template('马王堆简帛手写单字图像识别系统.html', img = img_stream)
@app.route('/马王堆医书长文本识别系统.html')
def index3():
    img_stream = return_img_stream("./static/马王堆医书长文本识别系统_files/default.png")
    return render_template('马王堆医书长文本识别系统.html', img = img_stream)
```

5.2.2 缺失文本预测模块功能实现

缺失文本预测模块的核心功能在前端主要包含一个文本输入框、一个提交按钮和一个文本区域。使用<form>表单和POST方法实现前端向后端提交数据请求的功能，当表单被提交时，文本输入框中的数据将作为名为“data”的参数被发送到服务器的/send_text路径，后端完成预测后，将预测结果作为“data_dict”返回给前端的文本区域“textBox”显示。在后端定义了一个处理POST请求的路由，当有一个向/send_text路由的POST请求时，调用send_string函数来处理该请求。send_string函数中使用Flask框架的request对象从POST请求中获取名为data的表单字段的值以接收用户提交的语料，接着调用predict_mask方法对用户提交的语料进行预测，其中predict_mask为封装好的缺失文本预测接口，最后使用render_template方法将预测结果data_dict返回给前端显示。具体代码如下：

```
#缺失文本预测接口
def predict_mask(model_path, input):
    unmasker = pipeline("fill-mask", model = model_path)
    result = ""
    for i in unmasker(input):
        result += "{'score':"+ str(i['score']) + ",'token_str':" + i['token_str'] + "}" +"\n"
    return result
```

```
@app.route('/send_text', methods=['POST'])
def send_string():
    data = request.form.get('data')
    result = predict_mask(data)
    return render_template('中医古籍缺失文本修复系统.html', data_dict=result,
data_dict1 = data,    status="error")
```

缺失文本预测模块页实现效果如图5.4所示，用户可以在输入框中输入以<mask>表示缺失文本的中医古籍语句，当用户点击提交按钮后，后端将接收用户输入的待预测的中医古籍语句，待后端预测完成后，右下侧文本域将显示后端返回的预测结果。

图5.4 缺失文本预测模块页

5.2.3 马王堆简帛手写单字识别模块功能实现

马王堆简帛手写单字识别模块的核心功能在前端主要包含一个文件输入框、一个提交按钮和一个文本区域以及一段JavaScript代码。JavaScript代码段用于监听文件输入框的变化，当用户选择一幅图像后，JavaScript将读取该图像并显示。和缺失文本预测模块类似，同样使用<form>表单和POST方法实现前端向后端提交数据请求的功能，当表单被提交时，文件输入框中的图像数据将作为名为“image”的参数被发送到服务器的/upload_word_picture路径，待后端成功接收图像并完成预测后，预测结果将作为“data_dict”返回给前端的文本区域“textBox”显示。马王堆简帛手写单字识别模块在后端定义了一个处理POST请求的路由，当有一个向/upload_word_picture路由的POST请求时，调用upload_word_picture函数来处理该请求。upload_word_picture函数中使用Flask框架的request对象从POST请求中获取名为image的表单字段值以接收用户上传的单字图像，接着调用predict_single_word方法对用户提交的单字图像进行预测，其中predict_single_word方法为封装好的手写单字识别接口，接口输入权重路径和待识别图像路径，输出为对待识别图像的识别结果，最后使用render_template方法将预测结果data_dict返回给前端显示。具体代码如下：

```
#马王堆简帛手写单字识别接口
def predict_single_word(weight_path,image_path):
    model = get_EfficientNet(3339)
    model.load_state_dict(torch.load(weight_path, map_location=torch.device('cpu')))
    f2 = open('./tools/label2class.json', 'r')
    info_data = json.load(f2)
    key_list = list(info_data.keys())
    val_list = list(info_data.values())
    image = PIL.Image.open(image_path)
    image = image.resize((64,64))
    transform = transforms.Compose([transforms.Resize((224,
224)),transforms.ToTensor(),transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))])
    image = transform(image)
    image = image.unsqueeze(0)
    model.eval()
    with torch.no_grad():
        outputs = model(image)
        _, predicted = torch.max(outputs.data, 1)
        return key_list[val_list.index(predicted.item())]
@app.route('/upload_word_picture', methods=['POST'])
def upload_word_picture():
    if 'image' not in request.files:
        return "No file part in the request", 400
    file = request.files['image']
    if file.filename == '':
        return "No selected file", 400
    if file:
        file.save("./static/upload/" + file.filename)
        label = "识别结果为 " + predict_single_word(weight_path,"./static/upload/" +
file.filename)
        img_stream = return_img_stream("./static/upload/" + file.filename)
        return render_template('马王堆简帛手写单字图像识别系统.html',
img=img_stream, data_dict=label, status="error")
```

马王堆简帛手写单字识别模块页如图5.5所示，用户可以在文件输入框中选择需要识别

的单字图像，文件输入框的右侧区域将会显示用户选择的图像，当用户点击提交按钮后，后端将接收用户上传的单字图像并进行预测，待预测完成后，右下侧文本域将显示后端返回的预测结果。

图5.5 马王堆简帛手写单字识别模块页

5.2.4 马王堆医书长文本序列识别模块功能实现

马王堆医书长文本序列识别模块的核心功能在前端主要包含一个文件输入框、一个提交按钮和一个文本区域以及一段JavaScript代码。和马王堆简帛手写单字识别模块功能类似，JavaScript代码段同样用于监听文件输入框的变化，当用户选择一个图像文件后，JavaScript将读取该图像并显示，同样使用<form>表单和POST方法实现前端向后端提交数据请求的功能。当表单被提交时，文件输入框中的图像数据将作为名为“image”的参数被发送到服务器的/upload_words_picture路径，待后端成功接收图像并完成预测后，前端的文本区域“textBox”将显示data_dict的值。马王堆医书长文本序列识别模块在后端定义了一个处理POST请求的路由，当有一个向/upload_words_picture路由的POST请求时，调用upload_words_picture函数来处理该请求。upload_words_picture函数中使用Flask框架的request对象从POST请求中获取名为image的表单字段值以接收用户上传的长文本序列图像，接着调用predict_words方法对用户提交的长文本序列图像进行预测，并使用correct_sentence方法对长本文识别模型的输出进行纠错，其中predict_words方法和correct_sentence方法分别为封装好的长文本识别接口和文本纠错接口，最后使用render_template方法将纠正后的预测结果data_dict返回给前端显示。具体代码如下：

```
#长文本识别接口
def predict_words(weight_path,image_path):
    alpha = cfg.word.get_all_words()
    net = crnn.CRNN(num_classes=len(alpha))
```

```
        net.load_state_dict(torch.load(weight_path, map_location='cpu')['model'])
        net.eval()
        image = load_image(image_path)
        image = torch.FloatTensor(image)
        predict = net(image)[0].detach().numpy()
        label = np.argmax(predict[:], axis=1)
        label = [alpha[class_id] for class_id in label]
        label = [k for k, g in itertools.groupby(list(label))]
        label = ' '.join(label)
    return label
    #文本纠错接口
    def correct_sentence(ckpt_path, vocab_path, cfg_path,input):
        tokenizer = BertTokenizerFast.from_pretrained(vocab_path)
        cfg.merge_from_file(cfg_path)
        model = MacBert4Csc.load_from_checkpoint(checkpoint_path=ckpt_path,
                                                 cfg=cfg,
                                                 map_location=device,
                                                 tokenizer=tokenizer)
        model.to(device)
        model.eval()
        inputs = tokenizer(input, padding=True, return_tensors='pt')
        inputs.to(cfg.MODEL.DEVICE)
        with torch.no_grad():
            outputs = model.forward(input)
            y_hat = torch.argmax(outputs[1], dim=-1)
            expand_text_lens = torch.sum(inputs['attention_mask'], dim=-1) - 1
        rst = []
        for t_len, _y_hat in zip(expand_text_lens, y_hat):
            rst.append(tokenizer.decode(_y_hat[1:t_len]).replace(' ', ''))
        return "".join(rst)
        @app.route('/upload_words_picture', methods=['POST'])
        def upload_words_picture():
            if 'image' not in request.files:
                return "No file part in the request", 400
```

```
        file = request.files['image']
        if file.filename == '':
            return "No selected file", 400
        if file:
            file.save("./static/upload/" + file.filename)
            label = predict_words(weight_path, "./static/upload/" + file.filename)
            correct_label = correct_sentence(ckpt_path, vocab_path, cfg_path, label)
            label = "识别结果为  " + correct_label
            img_stream = return_img_stream("./static/upload/" + file.filename)
            return render_template('马王堆医书长文本识别系
    统.html',img=img_stream, data_dict=label, status="error")
```

马王堆医书长文本序列识别模块页如图5.6所示，用户可以在文件输入框中选择需要识别的长文本序列图像，文件输入框的右侧区域将会显示用户选择的图像，当用户点击提交按钮后，后端将接收用户上传的序列图像并进行预测，待预测完成后，右下侧文本域将显示后端返回的预测结果。

图5.6 马王堆医书长文本序列识别模块页

5.3 本章小结

本章是关于中医古籍缺失文本预测和手写文字识别系统的设计和实现，首先对系统需求进行了分析，介绍了系统三大功能模块基本需求。接着介绍了系统总体操作流程，最后介绍了各功能模块的实现过程以及界面展示。

第六章 总结与展望

6.1 总结

中医古籍是中医药学传承的重要载体，也是中医药学术进步和创新的源泉。古籍是重要文献资源，是人类文明走向未来的文化物质基础。学界对于古籍价值的判定多依据古籍善本“三性”，即历史文物性、学术资料性及艺术代表性，认为古籍具有文物价值、学术价值与艺术价值。中医古籍是中医药知识传承的重要载体，是我国古籍的重要组成部分，见证中医学术进步与历史发展，蕴涵着丰富而宝贵的理论知识，承载着历代医家临证实践经验，是中医学术传承和创新的根基。本文将深度学习技术应用于中医古籍修复中缺失文本的预测以及于中医古籍的简帛手写文字图像识别中，期望为中医古籍研究者提供帮助，降低中医古籍文字的阅读难度。

本文的主要工作如下：

（1）收集并整理了一份83 032 043字的中医古籍语料库。该语料库收录了自秦汉以来的包括《黄帝内经》《金匮要略》《伤寒杂病论》《神农本草经》等经典中医古籍在内的608本中医古籍。

（2）创建了一份3 339类，每一类约3 000个样本的马王堆简帛手写文字平衡图像数据集。当前含有比较多的手写英文数据集，公开的手写中文数据集比较稀少，而公开的中医古籍手写文字图像数据集则更加稀少。《马王堆汉墓简帛文字全编》对马王堆所有简帛文字原始资料作了通盘的清理，秉持“有字必录”原则，收录了马王堆汉墓简帛中出现的3 339类文字。本文以《马王堆汉墓简帛文字全编》为基础，使用目标检测和分割算法，自动分割出《马王堆汉墓简帛文字全编》中所列出的单字图像，然后人工进行分类，形成原始单字图像数据集。而原始的单字图像数据集存在极大的数据不平衡性，因此论文又使用图像增强技术和生成对抗网络技术对原始图像数据集进行扩充，最终形成了平衡的马王堆简帛手写文字平衡图像数据集。同时在此数据集上训练了ResNet图像分类网络，验证该平衡数据集的有效性。

（3）将语言模型技术应用于中医古籍缺失文本的预测。构建了ACMBBERT模型，同时分别训练和测试N-gram模型、LSTM模型、BiLSTM模型、BERT模型，验证ACMBBERT在中缺失文本预测任务上的优势，并将ACMBBERT模型运用到文本修复场景中。实验结果表明，BiLSTM模型优于LSTM模型，LSTM模型明显优于N-gram模型，而BERT模型效果最优。在中医古籍-S语料库上训练的ACMBBERT-S模型达到了63.36%的hit@1，82.57%的hit@5。即ACMBBERT-S模型输出的前10条预测中，可以直接获得正确的缺失内容的几率为60.79%，出现正确预测的几率为82.57%。而在中医古籍-MWD语料库上训练的ACMBBERT-MWD模型达到了37.58%的hit@1，51.42%的hit@5和57.44%的hit@10，证明ACMBBERT模型在中医古籍缺失文本预测的任务上具备一定的泛化能力。

（4）提出了一种结合语言模型纠错的马王堆医书长文本识别模型。基于马王堆简帛平衡单字图像数据集，从每一类中随机抽取一张图片，拼接成随机长度的文本列，构建长文本数据集，作为长文本识别模型的训练数据。如此，可以生成大量长度不一的训练数据，无需从马王堆简帛原始图像中截取文本列，再不断地人工查阅《马王堆汉墓简帛文字全编》添加标签，极大提升了数据集构建的效率。基于生成的长文本数据集和CRNN算法，训练马王堆简帛长文本识别模型。该模型在生成数据集上识别准确率较高，达86.73%，而在真实的马王堆医书简帛图像上准确率不高，仅有63%。结合在ACMBBERT模型和马王堆医书文本语料库基础上微调来的马王堆医书文本纠错模型，马王堆医书简帛长文本识别模型的识别准确率得到了较大提升，达到了91%。同时，对于《天回医简》中截取的包含25幅单字图像和25幅长文本序列图像的测试集，马王堆简帛长文本识别模型能够准确地识别出其中19幅单字图像，虽然仅能完全准确地识别出4幅长文本序列图像，但是只有一个字识别错误的有12幅，同时25幅长文本序列图像共计134个字，模型准确识别的总字数为87，在长文本图像的字符维度上达到了60.44%的准确率。

（5）设计和实现了一个中医古籍缺失文本预测和简帛手写文字识别系统。系统支持用户输入隐去了一个或多个字的中医古籍文本，并输出缺失文本预测模型的预测。也支持用户上传马王堆简帛中的某一个文字图像或马王堆医书简帛的文本列图像，快速获取图像中文字内容，从而降低马王堆简帛的阅读门槛。

6.2 展望

尽管论文针对中医古籍智能修复做了大量的工作和研究，但是由于研究尚不成熟以及个人能力还有所欠缺等原因，本文还存在一些不足和需要改进的地方，最主要的是以下几点：

（1）尽管本文将语言模型应用于中医古籍缺失文本的预测中，并且训练了一个可以支持下游任务的中医古籍领域预训练模型，但是在模型预测准确率和语料库规模上仍有提升空间。

（2）尽管本文实现了对马王堆医书简帛长文本的识别，但是在实际应用中还需用户自行对简帛图像进行截取，作为文本识别模型的输入。后续，可以考虑构建马王堆医书简帛文本检测模型，对整幅简帛图像中的文本进行检测并分割，将分割出的文本列图像输入到结合文本纠错模型的马王堆医书文本识别模型中，从而实现端到端的文本识别。届时，用户只需提交一幅完整的马王堆医书的简帛图像，就可以自动获取图像中完整的纯文字信息，这将极大地推动马王堆医书的数字化进程，降低马王堆医书简帛的阅读难度，从而推动中医文化的传承与发展。

（3）中医古籍博大精深，中医古籍的修复和数字化是一项复杂、专业且精细的工作，深度学习技术可以应用于其中的很多环节。由于工作量的关系，本文只做到了将深度学习

技术应用于古籍缺失文本的预测和中医古籍简帛图像的文字识别，中医古籍研究中其他可以使用深度学习技术以提升中医古籍研究者工作效率的研究点仍有待发掘，例如古籍图像碎片的自动拼接，古籍图像缺失内容的自动填充等。

参考文献

[1] 李兵, 刘国正, 符永驰, 等. 从中医古籍数据库建设看中医古籍数字化[J]. 中国中医药信息杂志, 2009, 16(3): 92-93.

[2] 邱玏. 中医古籍英译历史的初步研究[D]. 北京: 中国中医科学院, 2011.

[3] 严顺. 先秦文献的语料库构建探究[J]. 江苏科技信息, 2016, 489(12): 32-33.

[4] 石民. 先秦汉语自动分词及词性标注研究[D]. 南京: 南京师范大学, 2010.

[5] 留金腾, 宋彦, 夏飞. 上古汉语分词及词性标注语料库的构建——以《淮南子》为范例[J]. 中文信息学报, 2013, 27(6):6-15,81.

[6] 石民, 李斌, 陈小荷.基于 CRF 的先秦汉语分词标注一体化研究[J].中文信息学报, 2010, 24(2): 39-45.

[7] 白玲玲. 中医药古文献语料库建设的语料分类问题研究[D]. 济南: 山东中医药大学, 2007.

[8] 付璐, 李思, 李明正, 等. 以清代医籍为例探讨中医古籍分词规范标准[J]. 中华中医药杂志, 2018, 33(10): 4700-4705.

[9] Assael Y, Sommerschield T, Shillingford B, et al. Restoring and attributing ancient texts using deep neural networks[J]. Nature, 2022, 603(7900): 280-283.

[10] Fetaya E, Lifshitz Y, Aaron E, et al. Restoration of fragmentary Babylonian texts using recurrent neural networks[J]. Proceedings of the National Academy of Sciences, 2020, 117(37): 22743-22751.

[11] Roark B, Saraclar M, Collins M. Discriminative n-gram language modeling[J]. Computer Speech & Language, 2007, 21(2): 373-392.

[12] Esmaeilzadeh A, Taghva K. Text classification using neural network language model (nnlm) and bert: An empirical comparison[C]//Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3. Springer International Publishing, 2022: 175-189.

[13] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[14] Meng K, Bau D, Andonian A, et al. Locating and editing factual associations in GPT[J]. Advances in Neural Information Processing Systems, 2022, 35: 17359-17372.

[15] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation[J]. arXiv preprint arXiv:1911.00536, 2019.

[16] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[17] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J].

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
[18] Huang K, Singh A, Chen S, et al. Clinical XLNet: modeling sequential clinical notes and predicting prolonged mechanical ventilation[J]. arXiv preprint arXiv:1912.11975, 2019.
[19] Liao M, Shi B, Bai X, et al. Textboxes: A fast text detector with a single deep neural network[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
[20] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676-3690.
[21] Zhou X, Yao C, Wen H, et al. East: an efficient and accurate scene text detector[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.
[22] Yuliang L, Lianwen J, Shuaitao Z, et al. Detecting curve text in the wild: New dataset and new solution[J]. arXiv preprint arXiv:1712.02170, 2017.
[23] Dai P, Zhang S, Zhang H, et al. Progressive contour regression for arbitrary-shape scene text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7393-7402.
[24] He M, Liao M, Yang Z, et al. MOST: A multi-oriented scene text detector with localization refinement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8813-8822.
[25] Wang Y, Xie H, Zha Z J, et al. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11753-11762.
[26] Zhang C, Liang B, Huang Z, et al. Look more than once: An accurate detector for text of arbitrary shapes[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10552-10561.
[27] Deng D, Liu H, Li X, et al. Pixellink: Detecting scene text via instance segmentation[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
[28] Wu Y, Natarajan P. Self-organized text detection with minimal post-processing via border learning[C]//proceedings of the IEEE international conference on computer vision. 2017: 5000-5009.
[29] Tian Z, Shu M, Lyu P, et al. Learning shape-aware embedding for scene text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4234-4243.
[30] Wang W, Xie E, Song X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//Proceedings of the IEEE/CVF international conference on

computer vision. 2019: 8440-8449.

[31] Liao M, Wan Z, Yao C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11474-11481.

[32] Zhu Y, Chen J, Liang L, et al. Fourier contour embedding for arbitrary-shaped text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3123-3131.

[33] Tang J, Yang Z, Wang Y, et al. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping[J]. Pattern recognition, 2019, 96: 106954.

[34] Xue C, Lu S, Zhang W. MSR: multi-scale shape regression for scene text detection[J]. arXiv preprint arXiv:1901.02596, 2019.

[35] Fu X, Chng E, Aickelin U, et al. CRNN: a joint neural network for redundancy detection[C]//2017 IEEE international conference on smart computing (SMARTCOMP). IEEE, 2017: 1-8.

[36] Xie Z, Huang Y, Zhu Y, et al. Aggregation cross-entropy for sequence recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6538-6547.

[37] Shi B, Wang X, Lyu P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4168-4176.

[38] Wang D, Liu C, Zhao Z, et al. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts[J]. arXiv preprint arXiv:2307.05354, 2023.

[39] Siriguleng. Automatic Punctuation Method of Ancient Chinese Texts Based on SikuBERT and Multi-head Attention Mechanism: An Exploration of Ancient Classical Ritual Literature[C]//2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2023, 11: 1255-1258.

[40] 沈思,陈猛,冯暑阳等.ChpoBERT:面向中文政策文本的预训练模型[J].情报学报, 2023, 42(12): 1487-1497.

[41] 刘欣华, 左丹妮, 范一丹等. 基于自然语言处理技术的互联网医疗眼表疾病咨询需求分析[J]. 中国卫生资源, 2023, 26(05): 527-533.

[42] 李利杰, 张君华. 基于循环神经网络和全局化领域的推荐算法[J]. 计算机与数字工程, 2022, 50(08): 1676-1679, 1701.

[43] 李健, 熊琦, 胡雅婷等. 基于 Transformer 和隐马尔科夫模型的中文命名实体识别方法

[J]. 吉林大学学报(工学版), 2023, 53(05): 1427-1434.
[44] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Interspeech. 2010, 2(3): 1045-1048.
[45] 孔丽雅, 周治平. 基于 ELMo 的混合注意力网络的方面级情感分析研究[J]. 中文信息学报, 2023, 37(06): 147-156.
[46] 王震宇, 朱学芳. 基于多模态 Transformer 的虚假新闻检测研究[J]. 情报学报, 2023, 42(12): 1477-1486.
[47] 杨冰, 那巍, 向学勤. 基于单阶段生成对抗网络的文本生成图像方法[J]. 浙江大学学报(工学版), 2023, 57(12): 2412-2420.
[48] Mariani G, Scheidegger F, Istrate R, et al. Bagan: Data augmentation with balancing gan[J]. arXiv preprint arXiv:1803.09655, 2018.
[49] Huang G, Jafari A H. Enhanced balancing GAN: Minority-class image generation[J]. Neural computing and applications, 2023, 35(7): 5145-5154.
[50] 裘锡圭. 长沙马王堆汉墓简帛集成[M]. 北京:中华书局, 2014.
[51] 刘钊, 郑健飞, 李霜洁, 等.马王堆汉墓简帛文字全编[M]. 北京: 中华书局, 2020.
[52] 周一谋, 萧佐桃. 马王堆医书考注[M]. 天津: 天津科学技术出版社, 1988.
[53] 严健民. 五十二病方注补译[M]. 北京: 中医古籍出版社,2005.
[54] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
[55] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[56] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[57] Sinha D, El-Sharkawy M. Thin mobilenet: An enhanced mobilenet architecture[C]//2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE, 2019: 0280-0285.
[58] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.
[59] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.
[60] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer

prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.

[61] Guan Y, Pang Z, Ding Y, et al. Text error correction after text recognition based on MacBERT4CSC[C]//Sixth International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2023). SPIE, 2023, 12787: 648-656.

[62] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[63] 天回医简整理组. 天回医简[M]. 北京: 文物出版社, 2022.

文献综述

基于深度学习的中医古籍修复研究综述

摘要：为了更好地了解深度学习技术及其在中医古籍修复领域的应用现状，为中医古籍修复的研究者提供参考，本文简要论述中医古籍修复领域的四个方向以及这四个方向的技术发展概况，分别为古汉语语料库构建研究现状、古籍缺失文本修复研究现状、古籍图像重建研究现状和古籍手写文字识别研究现状，从文本和图像双角度阐述了深度学习技术在中医古籍修复领域的可行性。基于深度学习进行中医古籍修复，一方面可以为中医古籍的研究者提供高质量的文本语料库，降低中医古籍修复专家和研究人员修复古医书的难度；另一方面可以让更多人了解并熟悉中医古籍中包含的中医、中药、养生和保健等知识，推广中医文化。

关键词：深度学习；文本修复；图像修复；中医古籍

Review of Research on Restoration of Ancient Chinese Medical Books Based on Deep Learning

Abstract: In order to better understand the deep learning technology and its application status in the field of traditional Chinese medicine ancient book restoration, and provide a reference for researchers in the field of traditional Chinese medicine ancient book restoration, this paper briefly discusses the three directions in the field of traditional Chinese medicine ancient book restoration and the technical development overview of the three directions, respectively, the research status of ancient Chinese corpus construction, the research status of ancient book missing text restoration, and the research status of ancient book image reconstruction, expounds the feasibility of deep learning technology in the field of ancient Chinese medicine book restoration from both text and image perspectives. The restoration of ancient Chinese medical books based on deep learning can, on the one hand, provide high-quality text corpus for researchers of ancient Chinese medical books, and reduce the difficulty of experts and researchers in the restoration of ancient Chinese medical books; On the other hand, it can let more people know and be familiar with the knowledge of traditional Chinese medicine, traditional Chinese medicine, health preservation and health care contained in ancient Chinese medical books, and promote the culture of traditional Chinese medicine.

Keywords：Deep learning; Text restoration; Image restoration; Ancient Chinese Medicine Books

引言

中医古籍是中医药学传承和发展的载体，既具有重要的学术价值，又具有相当的文物

价值，中医药古籍文献中所收载的理法、方药、养生保健知识是取之不尽、用之不竭的宝库，具有极高的实用价值[1]。中医古籍反映了中医药学在历史上为中华民族的繁衍昌盛作出的伟大贡献和取得的辉煌成就，更重要的是，它所蕴藏保存下来的民族文化遗产和智慧结晶必将在新的历史时期为人类医药科学事业的发展作出不可估量的贡献[2]。

但是，中医古籍在长期的流传过程中，不仅遭受自然的侵蚀，也会受到人为的损坏，同时中医专业图书馆的古籍藏书原本可能就没有得到很好的保存，即使经过修复，仍然存在古籍破损的现象，此外图书馆的书库环境、保护手段都可能造成中医古籍的损坏。因此针对中医古籍的整理和修复工作是十分必要的，而目前的修复工作主要是由古籍修复专家通过专业手段手动修复，但是这个过程需要对古汉语和中医古籍有广泛且深入的研究，是非常耗时和费力的。

由于中医古籍的专业性和特殊性，它们兼具古汉语和医学文献的特点，分析和处理难度较大。当前，已有一些学者和研究人员运用人工智能对中医古籍开展分析和研究，诞生了一些具有学术价值和应用价值的成果。

1 古汉语语料库构建研究现状

古汉语信息处理是当前中文信息处理领域一个重要研究方向，对古籍文献进行数字化和信息化的首要任务就是构建古汉语语料库(Corpus)。文献[3]对先秦文献的语料库构建进行了初步的探究，阐述了古文献语料库构建过程中涉及的古汉语分词以及词性标注等基本知识。在构建古籍语料库时需要制定对应的分词规范和标注集，南京师范大学语言科技研究所制定了先秦汉语词类标注基本集并构建了先秦典籍人工语料库[4]。文献[5]以《淮南子》为例介绍了上古汉语分词及词性标注语料库及其构建过程。文献[6]运用条件随机场方法对先秦汉语的分词和标注一体化工作开展研究。这些研究工作都没有考虑到中医古籍的特点，因此，在构建中医古籍语料库时，需要充分融入中医古籍内在特点对相应的方法进行合理的优化和改进。文献[7]对构建中医药古文献语料库的意义和思路进行了阐述，并且对语料分类问题进行了详细的介绍。文献[8]对中医古籍的分词规范进行了探索性的研究和总结，提出中医古籍分词规范建议，并以人工标注方式构建了一个小型的清代医籍分词语料库，但未能实现计算机自动分词，且语料库规模较小，不能推广和应用。

2 古籍缺失文本修复研究现状

由于年代久远，许多出土帛书残损非常严重，有诸多帛书碎片缺失，导致古医书文字残缺不全。如何对这些残字进行考释和修复是中医古籍研究人员面临的一个非常重要的问题。近年来，国外有学者开始尝试将深度学习技术应用于古文本的修复工作并取得了较好的效果。2019年，牛津大学与Google旗下国际顶尖的人工智能(Artificial-Intelligence, AI)企业DeepMind合作开展古文修复研究，联合开发了一款名为 Pythia的AI工具，利用深度学习

神经网络技术来识别和预测古希腊碑文中缺少的单词或字符，取得了较好的效果[9]。2020年9月，国际顶级期刊PNAS发表了几位以色列学者运用深度学习中的循环神经网络(Recurrent Neural Networks, RNN)来恢复残缺的古巴比伦文本的一项研究成果，实验结果表明该模型在预测残缺文本时效果良好，可以作为学者进行文本恢复工作的有力工具[10]。

如果将自然语言视为一系列离散标记$x_1, \ldots, x_T$,那么语言模型的任务就是要给出一个可以表示任意一个句子或序列出现的概率分布$p(x_1, \ldots, x_T)$。利用贝叶斯公式，可以将$p(x_1, \ldots, x_T)$分解为$\prod_{t=1}^{T} p$（$x_t|x_1, \ldots, x_{t-1}$）[10]，也就是给定前面的词，求后面一个词出现的条件概率。而这就意味着可以将中医古籍缺失文本修复的问题，简化为根据缺失内容的上下文信息来预测缺失内容的问题。模型经过训练后，向模型输入一条以“<mask>”标记缺字的中医古籍文本，模型将根据“<mask>”的上下文信息输出对“<mask>”的多个预测，例如输入《黄帝内经》中的“岐伯对曰：东方生<mask>，风生木，木生酸，酸生肝，肝生筋，筋生心，肝主目。其在天为玄，在人为道，在地为化。”，模型输出“风”、“木”、“酸”等多个输出，以及此处的<mask>为“风”、“木”、“酸”的概率。

常用的语言模型分为以N-gram[11]模型为代表的传统统计语言模型和以NNLM[12] (Neural Net Language Model)、长短期记忆 (Long Short-Term Memory，LSTM)模型[13]为代表的神经网络语言模型，以及后续由神经网络语言模型发展来的以BERT[14] (Bidirectional Encoder Representation from Transformers)为代表的预训练语言模型，如GPT[15](Generative Pre-Training) 、 RoBERTa[16](A Robustly Optimized BERT Pre-training Approach) 、XLNet[17](Generalized autoregressive pretraining for language understanding）等一系列模型。其中RoBERTa模型在BERT模型的基础上增加了大量训练参数和训练数据，在训练过程中使用更大的单次训练样本数，且在语言表征中使用了双字节编码，提高了词汇表征的准确度和任务执行效率[18]。预训练语言模型相较于传统的统计语言模型和神经网络语言模型，在大规模语料的处理上更具优势，预训练提供的模型初始化参数，可以使模型在目标任务上有更好的泛化性能和更快的收敛速度，同时在小规模数据上也能避免过拟合现象的发生。虽然预训练语言模型取得了很大成功，但是目前大多数预训练模型都是基于大量通用语料训练的，在面对特定领域文本的自然语言处理任务时，其功能的发挥容易受限。由于古代汉语在语法、语义、语用上与现代汉语存在较大差异，即使是面向中文构建的Chinese-BERT，在古汉语处理上也难以达到其在中文通用语料上的性能，直接调用预训练模型预测中医古籍中缺失的文本，并不能取得较好的效果[19]。因此需要构建高质量的中医古籍语料库，在中医古籍语料库的基础上训练面向中医古籍文本处理的预训练语言模型，最终训练的中医古籍预训练模型不仅可以用于中医古籍缺失内容的预测，同时也可以在微调之后用于中医古籍文本的命名实体识别、关系抽取、文本分类等下游任务。

3 古籍图像重建研究现状

图像重建任务中的碎片图像重组难题是计算机视觉和计算几何领域的一个经典问题，其主要目的是从给定数量且随机打乱的破碎图片中重建原始图像。图像重建研究技术的发展主要经历了四个历程，分别为使用形状信息进行图像重建、使用颜色信息进行图像重建、使用多重融合信息进行图像重建和近年来兴起的使用深层神经网络进行图像重建。

图像重建领域研究早期，学者们通过碎片的形状信息进行图像重建[20-31]，基于形状、轮廓、焦点、像素点等信息计算碎片之间的匹配关系。但是存在某些碎片形状匹配而边缘颜色纹理不匹配的情况，此时需要利用颜色信息来完成图像的重建[32-34]。基于相邻碎片图像的边缘区域颜色信息是相似的，可以利用边缘区域颜色的过渡平滑程度，设定一个阈值，判断两块碎片是否是相邻的。随着研究深入，同时考虑碎片的形状和颜色信息可以提高重建结果，也能够增强模型的鲁棒性[35-41]。但是，无论是基于颜色信息、基于形状信息，还是将颜色信息和形状信息融合，构建出来的算法往往只适用于某一套图片，缺乏较强的迁移性，随着深度学习技术的兴起，基于卷积神经网络强大的特征提取能力[42]，可以很好地提取碎片图像的各种特征，近年来出现了一些使用深度学习技术完成碎片图像的重建的研究[43-47]，其中一个名为JigsawNet的CNN模型，适配了碎片数量大，碎片形状随机的复杂问题，具有很强的鲁棒性。

4 古籍手写文字识别研究现状

古籍是研究历史文化，挖掘古人智慧的重要载体。针对古籍的文字识别，是古籍数字化的一大重要研究问题。文字识别属于光学字符识别的范畴，但是古籍的文字识别又存在其特殊性。古籍中的古汉字写法，在历朝历代都存在差异性，其次存在大量的生僻字，数据有着很大的不平衡性。此外，古籍由于年代久远等问题，图像不清晰，甚至有破损也是一大问题。古籍文字识别分为两个部分，首相是文字检测，然后是对检测到的文字进行识别或者分类。目前比较流行的文字检测的方法有基于回归的文本检测方法[48-55]和基于分割的文本检测方法[56-64]。基于回归的方法在水平文本检测有着不错的表现，但是对于弯曲文本效果不佳。而基于分割的方法先从像素层面做分类，判别每一个像素点是否属于一个文本目标，得到文本区域的概率图，再通过后处理的方式得到文本分割区域的包围曲线，此类方法对不规则形状的文本检测具有天然优势。文本识别方法主要分为规则文本识别[65-68]和不规则文本识别[69-77]。规则文本识别又分为基于 CTC（Conectionist Temporal Classification）的算法和基于 Sequence2Sequence 的算法，两种算法在规则文本识别领域都被证明是有效的。而古籍的文本通常是竖直排列的，相对来说比较规则，因此现有的文本检测和识别方法可以应用于古籍的文本识别，同时也需要考虑古籍文本的特殊性。

5 结语

纵观前人的研究成果，针对基于深度学习修复中医古籍，未来的研究方向可向以下四

个方向聚焦。第一，构建高质量的中医古籍语料库；第二，从文本角度运用语言模型技术对中医古籍中的缺失文本进行填补修复，同时可以基于构建的高质量的中医古籍语料库训练预训练模型，最终训练的中医古籍预训练模型不仅可以用于中医古籍缺失内容的预测，也可以在微调之后用于中医古籍文本的命名实体识别、关系抽取、文本分类等下游任务；第三，从图像角度，使用卷积神经网络系列技术实现中医古籍碎片图像的自动拼接；第四，从古籍手写文字识别角度，使用通用文本检测和识别框架，微调进行中医古籍领域适配，实现对中医古籍的检测与识别，以及识别古籍中部分比划残缺或笔迹模糊的文字，从而推动中医古籍的数字化、信息化。

总之，运用人工智能技术，特别是近年来发展迅速的深度学习技术，从中医古籍语料库的构建、古医书的修复等方面开展研究工作，深入挖掘和萃取中医古籍的精华，凝练和提取中医古籍所蕴含的经验和智慧，可以有效推动中医古籍和中医文化研究的信息化和智能化。一方面可以为中医古籍的研究者提供高质量的文本语料库，降低中医古籍修复专家和研究人员修复古医书的难度；另一方面可以让更多人了解并熟悉中医古籍中包含的中医、中药、养生和保健等知识，推广中医文化。

参考文献

[1] 李兵, 刘国正, 符永驰. 从中医古籍数据库建设看中医古籍数字化[J]. 中国中医药信息杂志, 2009, 16(3): 92-93.

[2] 邱玏. 中医古籍英译历史的初步研究[D]. 北京: 中国中医科学院, 2011.

[3] 严顺. 先秦文献的语料库构建探究[J]. 江苏科技信息, 2016, 489(12): 32-33.

[4] 石民. 先秦汉语自动分词及词性标注研究[D]. 南京: 南京师范大学, 2010.

[5] 留金腾, 宋彦, 夏飞. 上古汉语分词及词性标注语料库的构建——以《淮南子》为范例[J]. 中文信息学报, 2013, 27(6): 6-15, 81.

[6] 石民, 李斌, 陈小荷. 基于CRF的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010, 24(2): 39-45.

[7] 白玲玲. 中医药古文献语料库建设的语料分类问题研究[D]. 济南: 山东中医药大学, 2007.

[8] 付璐, 李思, 李明正, 等. 以清代医籍为例探讨中医古籍分词规范标准[J]. 中华中医药杂志, 2018, 33(10): 4700-4705.

[9] Assael Y, Sommerschield T, Shillingford B, et al. Restoring and attributing ancient texts using deep neural networks[J]. Nature, 2022, 603(7900): 280-283.

[10] Fetaya E, Lifshitz Y, Aaron E, et al. Restoration of fragmentary Babylonian texts using recurrent neural networks[J]. Proceedings of the National Academy of Sciences, 2020, 117(37): 22743-22751.

[11] Roark B, Saraclar M, Collins M. Discriminative n-gram language modeling[J]. Computer Speech & Language, 2007, 21(2): 373-392.

[12] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[13] 高佳奕, 杨涛, 董海艳等. 基于LSTM-CRF的中医医案症状命名实体抽取研究[J]. 中国中医药信息杂志, 2021, 28(05): 20-24.

[14] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the Association for Computational Linguistics. Stroudsburg, 2019: 4171-4186.

[15] 侯丹阳, 庞亮, 丁汉星等. 语言模型攻击性的自动评价方法[J]. 中文信息学报, 2022, 36(01): 12-20.

[16] 陆泉, 郝志同, 陈静等. 利用迁移学习精准识别领域信息之探讨[J]. 图书情报工作, 2021, 65(05): 110-117.

[17] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for language understanding[C]//Proceedings of Advances in Neural Information Processing Systems, 2019: 5753-5763.

[18] 王华锋, 王久阳. 一种基于Roberta的中文实体关系联合抽取模型[J]. 北方工业大学学报, 2020, 32(02): 90-98.

[19] 王东波, 刘畅, 朱子赫等. SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛, 2022, 42(06): 31-43.

[20] Radack G M, Badler N I. Jigsaw puzzle matching using a boundary-centered polar encoding[J]. Computer Graphics and Image Processing, 1982, 19(1): 1-17.

[21] Papaodysseus C, Panagopoulos T, Exarhos M, et al. Contour-shape based reconstruction of fragmented, 1600 bc wall paintings[J]. IEEE Transactions on Signal Processing, 2002, 50(6): 1277-1288.

[22] 贾海燕, 朱良家, 周宗潭, 等.一种碎纸自动拼接中的形状匹配方法[J]. 计算机仿真, 2006(11): 180-183.

[23] 张欣, 卜彦龙, 朱良家, 等.物证复原系统中的碎纸轮廓提取技术研究[J]. 计算机仿真, 2006(11): 184-187, 279.

[24] 蓝洋, 和亮. 基于0-1规划的规则中文文件碎片自动拼接技术[J]. 计算机系统应用, 2015, 24(04): 270-273.

[25] 黄薇, 许勇, 刘淑影. 基于像素点的碎片图像拼接技术[J]. 计算机应用, 2014, 34(S2): 269-270+290.

[26] 郑蓓蓓, 郭立本. 改进的遗传算法应用于碎片拼接[J]. 计算机与现代化, 2011(05): 52-

56.

[27] 牟廉明, 罗开宝, 陈琳, 等. 图像碎纸片半自动可视化拼接方法[J]. 计算机工程与应用, 2016, 52(01): 206-209, 238.

[28] 李旭茹, 韩冬, 刘德磊. 一种基于形状匹配的碎纸片拼接系统设计[J]. 电脑知识与技术, 2017, 13(16): 148-150.

[29] 廖海波, 李萍. 一种基于轮廓角点与灰度的二维碎片拼接[J]. 激光杂志, 2014, 35(12): 21-24.

[30] 沈鸿平, 章毅鹏, 王义康. 基于0-1规划模型的规则中文碎片拼接复原研究[J]. 电子科技, 2014, 27(06): 13-16, 21.

[31] Son K, Hays J, Cooper D B. Solving small-piece jigsaw puzzles by growing consensus[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1193-1201.

[32] Amigoni F, Gazzani S, Podico S. A method for reassembling fragments in image reconstruction[C]//Proceedings 2003 International Conference on Image Processing (Cat.No. 03CH37429). IEEE, 2003, 3: III-581.

[33] Tsamoura E, Pitas I. Automatic color based reassembly of fragmented images and paintings[J]. IEEE Transactions on Image Processing, 2009, 19(3): 680-690.

[34] Sagiroglu M S, Erçil A. A texture based matching approach for automated assembly of puzzles[C]//18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 3: 1036-1041.

[35] Cao S, Liu H, Yan S. Automated assembly of shredded pieces from multiple photos[C]//2010 IEEE International Conference on Multimedia and Expo. IEEE, 2010: 358-363.

[36] Cho T S, Avidan S, Freeman W T. A probabilistic image jigsaw puzzle solver[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 183-190.

[37] Pomeranz D, Shemesh M, Ben-Shahar O. A fully automated greedy square jigsaw puzzle database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[38] Liu H, Cao S, Yan S. Automated assembly of shredded pieces from multiple photos[J]. IEEE transactions on multimedia, 2011, 13(5): 1154-1162.

[39] Gallagher A C. Jigsaw puzzles with pieces of unknown orientation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 382-389.

[40] Richter F, Ries C X, Cebron N, et al. Learning to reassemble shredded documents[J]. IEEE Transactions on multimedia, 2012, 15(3): 582-593.

[41] Li X, Xie K, Hong W, et al. Hierarchical fragmented image reassembly using a bundle-of-superpixel representation[J]. Computer Aided Geometric Design, 2019, 71: 220-230.

[42] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee,2009: 248-255.

[43] Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]//European Conference on Computer Vision. Springer, Cham, 2016: 69-84.

[44] Kim D, Cho D, Yoo D, et al. Learning image representations by completing damaged jigsaw puzzles[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 793-802.

[45] Paumard M M, Picard D, Tabia H. Jigsaw puzzle solving using local feature co-occurrences in deep neural networks[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 1018-1022.

[46] Carlucci F M, D'Innocente A, Bucci S, et al. Domain generalization by solving jigsaw puzzles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2229-2238.

[47] Le C, Li X. JigsawNet: Shredded image reassembly using convolutional neural network and loop-based composition[J]. IEEE Transactions on Image Processing, 2019, 28(8): 4000-4015.

[48] Liao M, Shi B, Bai X, et al. Textboxes: A fast text detector with a single deep neural network[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).

[49] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676-3690.

[50]. Zhou X, Yao C, Wen H, et al. East: an efficient and accurate scene text detector[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.

[51] Yuliang L, Lianwen J, Shuaitao Z, et al. Detecting curve text in the wild: New dataset and new solution[J]. arXiv preprint arXiv:1712.02170, 2017.

[52] Dai P, Zhang S, Zhang H, et al. Progressive contour regression for arbitrary-shape scene text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7393-7402.

[53] He M, Liao M, Yang Z, et al. MOST: A multi-oriented scene text detector with localization refinement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8813-8822.

[54] Wang Y, Xie H, Zha Z J, et al. Contournet: Taking a further step toward accurate arbitrary-

shaped scene text detection[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11753-11762.

[55] Zhang C, Liang B, Huang Z, et al. Look more than once: An accurate detector for text of arbitrary shapes[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10552-10561.

[56] Deng D, Liu H, Li X, et al. Pixellink: Detecting scene text via instance segmentation[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[57] Wu Y, Natarajan P. Self-organized text detection with minimal post-processing via border learning[C]//proceedings of the IEEE international conference on computer vision. 2017: 5000-5009.

[59] Tian Z, Shu M, Lyu P, et al. Learning shape-aware embedding for scene text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4234-4243.

[60] Wang W, Xie E, Song X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8440-8449.

[61] Liao M, Wan Z, Yao C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11474-11481.

[62] Zhu Y, Chen J, Liang L, et al. Fourier contour embedding for arbitrary-shaped text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3123-3131.

[63] Tang J, Yang Z, Wang Y, et al. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping[J]. Pattern recognition, 2019, 96: 106954.

[64] Xue C, Lu S, Zhang W. MSR: multi-scale shape regression for scene text detection[J]. arXiv preprint arXiv:1901.02596, 2019.

[65] 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, 2016, 39(11): 2298-2304.

[66] Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 71-79.

[67] Gao Y, Chen Y, Wang J, et al. Reading scene text with attention convolutional sequence modeling[J]. arXiv preprint arXiv:1709.04303, 2017.

[68] Shi B, Wang X, Lyu P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4168-4176.

[69] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.

[70] 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.

[71] Lee C Y, Osindero S. Recursive recurrent nets with attention modeling for ocr in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2231-2239.

[72] Li H, Wang P, Shen C, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8610-8617.

[73] Lyu P, Yao C, Wu W, et al. Multi-oriented scene text detection via corner localization and region segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7553-7563.

[74] Liao M, Zhang J, Wan Z, et al. Scene text recognition from two-dimensional perspective[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8714-8721.

[75] Yu D, Li X, Zhang C, et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12113-12122.

[76] Sheng F, Chen Z, Xu B. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition[C]//2019 International conference on document analysis and recognition (ICDAR). IEEE, 2019: 781-786.

[77] Yang L, Wang P, Li H, et al. A holistic representation guided attention network for scene text recognition[J]. Neurocomputing, 2020, 414: 67-75.