



桂林电子科技大学
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

本科毕业设计（论文）

题目： 基于深度学习的手写报告智能评分系统

学 号： 1800800222

姓 名： 刘舒豪

学 院： 信息与通信学院

专 业： 通信工程

指 导 教 师： 首照宇

指导教师职称： 研究员

2022 年 5 月 20 日

摘 要

文本检测识别和自然语言处理是计算机应用领域中两个极其重要的研究方向。近年来基于深度学习的方法在这两个领域的任务中表现格外出色，相比于传统的处理方法有着巨大的优势。然而，在当前面临的文本检测识别任务中，对于手写文本字符的识别仍然面临着许多的困难，对于模糊手写文本和打印字体与手写字体混用的实际场景识别和检测效果性能较差。同时，在自然语言处理任务中，对于文章的量化评分方法也存在着许多不足之处。因此，本文通过深度学习的方法进一步提高模型对于混用文本的检测识别能力，并将手写字符文本检测识别与报告评分应用场景结合在一起。本文的研究内容如下：

- (1) 通过搜集资料了解并学习了目前文本检测与识别和自然语言处理的前端技术，通过检测识别对比测试和模块化整合测试，搭建了基于 `paddle` 的结合目标检测和目标分割的 OCR 模型，并基于预训练模型进行训练优化。
- (2) 调整参数基于 `paddle` 工具整合 CRNN 网络成功训练了对于手写字符文本识别率较高的文本识别处理模型，并成功完成了中科院自动化研究所手写中文数据集的评估检验，得出了较好的识别结果。
- (3) 调试并测试了基于 Bert 的语言表征模型的 Bleurt 自然语言处理评分模型，为了模块化处理设计了分类处理版面文字的算法并进行模块化输出，同时将其与文本检测识别模块相对接处理。
- (4) 搭建了面向用户和教师端使用的服务器系统，设计并实现了具有图片上传、预览功能的前端网页页面，同时设计并实现了服务器端接收和存储的功能，连接实现了服务器端与代码保存管理工具的对接功能。

关键词：深度学习；文本检测识别；自然语言处理；卷积神经网络

Abstract

Text detection and recognition and natural language processing are two extremely important research directions in the field of computer applications. In recent years, deep learning-based methods have performed exceptionally well in these two areas, with great advantages over traditional processing methods. However, in current text detection and recognition tasks, the recognition of handwritten text characters still has many difficulties, and the recognition with detection performance is poor for the actual scenarios where fuzzy handwritten text and print fonts are mixed with handwritten fonts. Also, there are many shortcomings in the quantitative scoring methods for articles in natural language processing tasks. Therefore, this paper further improves the model's detection and recognition ability for mixed text through a deep learning approach and combines handwritten character text detection and recognition with report scoring application scenarios. The research in this paper is as follows:

(1) Understand and learned the current front-end technologies of text detection and recognition and natural language processing by collecting information. Built a paddle-based OCR model combining target detection and target segmentation through detection and recognition comparison tests and modular integration tests, and conducted training optimization based on the pre-trained model.

(2) Adjusted parameters based on paddle tool to integrate CRNN network successfully trained a text recognition processing model with high recognition rate for handwritten character text, and successfully completed the evaluation test of handwritten Chinese dataset from the Institute of Automation, Chinese Academy of Sciences, and obtained better recognition results.

(3) Adjusted and tested Bleurt's natural language processing scoring model based on Bert's language representation model, designed an algorithm to classify and process layout text for modular processing, and processed it relative to the text detection and recognition module.

(4) Built a server system for users and teachers. Designed and implemented a front-end web page with image upload and preview functions. Designed and implemented server-side receiving and storage functions. At the same time connected the server-side to the code saving management tool.

Key words: Deep Learning; OCR; NLP; CNN

目 录

引言	1
1 绪论	1
2 深度学习理论基础	2
2.1 文本检测	3
2.1.1 基于目标检测的方法	3
2.1.2 基于目标分割的方法	3
2.2 文字字符识别	3
2.2.1 CRNN(卷积递归神经网络)技术背景	3
2.2.2 卷积递归神经网络的特征序列提取方法	4
2.2.3 卷积递归神经网络的翻译	5
2.2.4 卷积神经网络基于词汇的转录	5
2.2.5 卷积神经网络的网络训练	5
2.2.6 CRNN 模型训练数据集	5
2.2.7 CRNN 模型总结	6
2.3 NLP 自然语言处理	6
3 文本检测与识别的具体实现过程	7
3.1 对于现有模型的应用测试	7
3.1.1 本次项目中所试用过的文本检测方法	7
3.1.2 本次项目中所试用过的文本识别方法	7
3.2 基于 Paddle 开发者工具实现手写报告文本检测与识别功能	8
3.2.1 文本检测	8
3.2.2 文本识别	8
3.3 文字检测与识别深度学习具体训练过程	8
3.3.1 数据集搜集与标签化整理	8
3.3.2 预训练模型	9
3.3.3 训练模型网络结构设计	9
3.3.4 训练结果与过程记录	10
4 基于 BLEURT 的 NLP 文本评分	15
5 基于服务器实现的网页形式前端客户端设计	17
6 结论	21
谢 辞	22
参考文献	23

引言

由于教育现代化的进一步发展，学生课业活动中所培养的能力更加多元化，为了提升学生的科研素质和科技创新能力，教学中课堂实验和自主实验的比例有了明显的增长，也正是因为如此，老师们对于大量的手写文字报告的批改与评分流程化便成为了需求。随着深度学习技术的进步，对于手写文字字符识别和自然语言处理有了较好的解决方法，将其应用于手写文字报告的批改与评分的方法能够明显节省教师们课后的实验报告批改时间，提升实验报告批改效率，从而使得教师能更好地进行教学备课活动，同时也能以更快更便捷的方法使得学生获得及时的报告批改反馈。在系统流程中，实现无纸化传输批改，减少了报告递交和批改过程中墨水和纸张的消耗，节约资源，保护了环境，有利于实现低碳化社会。

1 绪论

近年来深度学习领域收到了许多的关注，尤其在光学字符识别和自然语言处理领域。

光学字符识别 (Optical character recognition)谷歌近年热点图如图 1-1 所示。

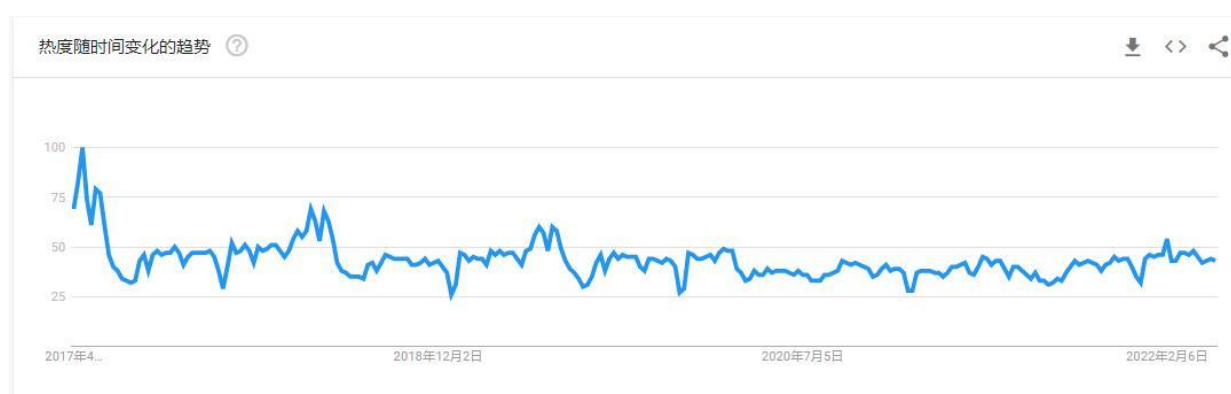


图 1-1 Google Trends 关于 OCR 搜索热度图

自然语言处理(Natural language processing)谷歌近年热点图如图 1-2 所示。



图 1-2 Google Trends 关于 NLP 搜索热度图

当前在相关领域内的研究成功实现了对于毛笔字迹的书法评价，对于英语作文的批改评分等项目，但是对于手写中文文字报告的智能评分方法的研究目前仍然较少。

当前文本检测识别研究搜索热度主要在英国，中国，斯洛伐克，尼泊尔等国家。当

前自然语言处理研究搜索热度主要在中国，以色列，荷兰，斯洛文尼亚等国家。中国国内的中文文本的检测识别模型当前主要由 Clue 团队进行评估与记录。

当前 CLUE1.1 阅读理解排行榜如图 1-3 所示。

排行	模型	研究机构	测评时间	Score1.1	认证	CMRC2018	CHID 1.1	C3 1.1
1	ShenNonG	云小微AI	21-12-01	91.734	待认证	86.500	97.932	90.769
2	ShenZhou	QQ浏览器实验室(QQ Browser Lab)	21-09-19	91.627	待认证	87.850	97.923	89.108
3	bert	antins	22-04-28	91.569	待认证	85.100	98.407	91.200
4	deberta-large	search-nlp	22-02-25	89.346	待认证	82.450	93.896	91.692
5	MRC	云小微	21-09-09	87.945	待认证	84.600	91.790	87.446
6	Mengzi	澜舟科技-创新工场	21-09-14	86.291	待认证	83.950	95.109	79.815
7	CL-BERT	CL-BERT	22-04-06	85.836	待认证	81.950	98.512	77.046
8	GEB	哥德尔	22-03-31	83.408	待认证	81.950	91.227	77.046
9	MMM	一个队伍的名字	22-03-31	83.408	待认证	81.950	91.227	77.046
10	Ensemble	Hunyuan 1.0	22-04-01	83.368	待认证	79.950	93.722	76.431
11	RoFormerV2 large	追一科技	22-03-19	83.181	待认证	80.500	90.705	78.338
12	RoBERTa large - baseline	bojone	21-11-12	82.683	待认证	78.850	92.460	76.738
13	BERT	BERT	21-10-18	82.572	待认证	70.000	89.901	87.815

图 1-3 CLUE1.1 截至 2022 年 4 月 28 日阅读理解排行榜

相关领域近期成果有：有史以来最为庞大的量子计算自然语言处理任务由 CQC 公司声称实现。“AI”辩论家系统由 IBM 团队进行发布，已经可以与人类专家选手进行逻辑清晰论题深刻的现场辩论，通过数据库自行组织开场白和进行论点的反驳。微软 Ravanelli 团队将纯 PyTorch 的语言工具包 Speech Brain 进行了开源。清华大学联合智源研究院研究论证了 GPT 的语言理解能力达到与 BERT 相近水平的性能。

本文研究基于深度学习的方法，将手写文字字符 OCR 算法识别和自然语言处理算法相结合，通过服务器网页前端实现的方式开发了一套手写报告智能评分系统。

2 深度学习理论基础

OCR（OpticalCharacterRecognition，光学字符识别）这个英文简称就是目前基于图像处理方法进行文字识别的总体称谓，它包括且不仅仅限制于书本，印刷文档识别，同时也包括在各种实际应用场景，各种生活环境下的文字字符识别。文字识别通常流程中包括两个部分，通过文本检测算法检测出所测图片中的文本行，再对所检测出的文本行内容通过识别算法识别出具体文字内容。NLP，中文名为自然语言处理。其是在人工智能和语言学研究方面的一个重要发展方向。它的重点研究内容就是分析如何处理自然语言并正确使用自然语言于目标场景中。计算机通过深度学习的技术对语言认知进行探索，同时追求对于文本内容的理解，在此基础上实现模拟自然语言的文本生成。自然语言认知与理解主要通过计算机将获取的语言转化为含有信息内容的符号与关系，然后再按照所设计目标的要求进行处理从而达到目的。自然语言生成系统则是为了将计算机数据经

过系统处理变为人类能够读取并且理解的常规语言形式。

2.1 文本检测

当前的文本检测模型主要分下面说的三大种，基于目标的检测方法、基于分割的检测方法以及综合目标检测和分割运用的方法。

2.1.1 基于目标检测的方法

通过预测得到文本框，再通过 NMS 筛选获取最终需要处理的内容，这些内容大多数是通过四点式文本框来框选的。这种方法的局限性主要在于只有在标准文字场景下才能较好地实现效果，而对于弯曲等非传统文本场景情况检测效果达不到理想情况。目前此类算法中的典型有 CRAFT、EAST、TextBox 等。

2.1.2 基于目标分割的方法

主要实现方法是将文本行作为分割目标，接着利用分割结果构建外接文本框的方法来实现文本检测。这种方法对于弯曲变形文本有较好的适应处理能力，但是面临文本交叉的情况时，其实现效果满足不了较高的检测要求。当前此类算法中的典型为 LSAE、DB、PSNet 等方法。

2.2 文字字符识别

当前的文本识别方法主要是通过对以文本检测方法生成的文本行信息进行识别处理，具体来说有基于 CTC 和基于 Attention 的主流文字预测方法，通过在卷积神经网络(CNN)+循环神经网络(RNN)后叠加文字预测方法来实现文本识别的功能。

2.2.1 CRNN(卷积递归神经网络)技术背景

随着神经网络的迅速发展，深度神经网络模型，特别是深度卷积网络在视觉任务实现了里程碑性的突破进展。由于多数与神经网络相关的工作都集中在对象类别的检测或者分类方法，基于图像的序列识别内的经典方向面临了对于突破当前限制而发展的新技术的需求。在现实生活环境中，大量的需求场景识别对象，比如自然场景文字、书写笔迹、乐谱等，这些对象往往以非孤立而连续序列的形式表现于环境中。由于这个缘故，在这类对象的识别中，往往都是通过将这些场景识别问题作为序列识别问题处理。这一类序列对象同时也具有另外一个长度不固定性的特征。根据^[1]中描述可知，而早期的深度模型并不能直接用于序列对象的检测，导致这类不适配性的主要原因在于过去的模型对于输入和输出具有固定维度的操作要求，并没办法生成可变长度的标签序列应用于模型处理与训练，由此可以发现此类系统并不能直接用于基于图像的序列识别。

本次 OCR 模型系统所应用的卷积神经网络架构是专门为识别图像中的序列状对象而设计的，其简称构成方式也表示了他的构成形式，也就是 DCNN 和 RNN 的组合应用。由^[2]中描述我们可以得知此模型对于类序列类型的对象识别处理与传统的神经网络模型相比具有多个显著的优势，这些优势如下：

(1) 卷积递归神经网络可以直接从序列标签而进行学习和训练，此过程中并不需要详细的注释内容。

(2) CRNN 与 DCNN 都是直接从图像数据中学习到信息显示含义,而并不需要人工对图像数据进行特征标注和进行预处理的措施,因此他们具有明显相同的特征性质。

(3) 产生标签序列这样的一致属性特性是 CRNN 与 RNN 所共有的。

(4) 卷积递归神经网络仅仅需要在训练和测试阶段中进行高度归一化处理即可,因为其并没对所处理的序列对象长度加以限制。

(5) 在场景文本特别是单词识别上卷积递归神经网络取得了比过去技术更加优越的性能。

(6) 卷积递归神经网络所包含的参数相比于传统模型,更加的精简化,所需要占用的存储空间也因此而更少。

2.2.2 卷积递归神经网络的特征序列提取方法

在本模块中,首先就必须设法完成特征序列捕捉并获取的功能。卷积层是由通过标准卷积神经网络模型中的筛选并抽离出卷积层和最大池化层所构建而成的。这样的组合方法的来由主要是运用在输出图像中抽取序列的特征来表示数据。并且为达到良好的处理效率,把全部的图像都缩放到了相同的高度,再分别进行在输入与输出网络之间的传输操作。这样在以后可以凭借卷积层组件所产生的特征图,对其进行有目的处理,从而得到特征向量序列,从而方便实现了循环层的输入过程。具体实现方法是通过将每列的宽度固定为单个像素,通过让特征序列的每个特征向量在特征图上按照列的形式从左到右依次生成的方法来实现目的,也就是将特征向量代表所有映射的相对对应串联内容。

过去落后的方法通过使用 CNN 来优化类似序列的对象的学习抗干扰能力,但是这些方法通常是利用 CNN 提取整体表示的图像,然后将局部深层特征信息提取后用来依次识别序列类对象的组件。在它们具有明显长度的状况下,因为需要 CNN 将输入图像缩放到固定的大小来满足其固定的输入尺寸,无法较好的具有此类特征性质的类似序列的对象。在不同类型的视觉识别任务中这些满足稳定、丰富和可训练的特征性质也已经被关注和设计。我们可以在 CRNN 中将这此深层特征归纳总结,并且可以将其传达到顺序表示中,由此可以保证对类似序列对象的长度变化保持连贯和一致性。

为了实现一个深度的神经网络构建于卷积层的顶部,使其满足双向循环特点,选择而利用序列标记的方法来处理所放置的循环层的架构问题。循环层提供特征序列的每个帧部分进行预测标签分布的功能。循环层在网络中具有不可忽视的重要作用是由于循环层具有多个方面的优点。第一,循环神经网络可以将误差差异反向传播至其前端,即卷积层,允许我们在一致的网络中将循环层和卷积层进行一体化的训练。第二,循环神经网络本身捕获序列中上下文所包含的信息的能力尤为出色。为了进行基于图像的序列识别通过使用上下文线索的方法会比较于单独处理每个符号更稳定和有用。循环神经网络能对任意序列进行操作,而不限其长度,具备灵活性,能够从头到尾实现遍历效果。

在传统的循环神经网络单元不可避免的面临一个问题,那就是梯度消失。参考^[3]可以看出,由于这个固定存在的问题,每次处理训练的过程中都无法进行广阔的区域数据

采集，导致它可以存储的上下文的范围收到了约束，同时加重训练过程中的负担，导致到达一定程度后学习能力效率产生了显然的下降，并且很难提升已有模型的识别准确性。为了解决这个问题，开发者们设计出了 Long-Short Term Memory 这样的循环神经网络单元。LSTM 包含了一个记忆单元和三个乘法门，即输入、输出和遗忘门，这样的网络单元可以实现图像序列之间的依赖特性，从而显著的提高了系统的性能。

2.2.3 卷积递归神经网络的翻译

开发者们提出在连接主义时态分类（CTC）层中定义的条件概率处理过程中，以帧预测的前提状况，将序列定义为概率。为了削弱忽略了序列中每个标签所在的位置这样操作导致的影响，面对负对数似然作为训练网络的目标时，为了满足要求使用该概率来实现直接优化从而免除了标记每个字符位置的过程，仅仅需要获取图像和它所对应的标签序列，并且据此设定了训练框架中的头部和 loss 等参数文件。

2.2.4 卷积递归神经网络基于词汇的转录

在卷积神经网络中，每个测试样本都同时与一个对应的词典相关联的特性是在基于词典的模式中所必然存在的特性。当面临的情况是样本与具有五十万字或者以上的词典对比时，对于词典进行详尽的搜索将耗费大量时间和可利用的硬件性能。卷积递归神经网络模型设计者为了应对为词典中的所有序列计算公式都处理并且得出一个概率，同时提取出概率最高的一个，在此过程中很容易错误或者数据损失的情况的问题，通过编组最近待选内容的搜索方法，研究得出通过无词典转录预测的标签序列经过特定方法进行量化分析的结果通常接近实际情况的值。

2.2.5 卷积递归神经网络的网络训练

当前存在利用训练图像，地面实况标签序列等方法来表示训练数据集。目标是最小化基本事实概率的负对数相似值。对成对的图像和序列进行处理时，便可以采取更为简化利用通过目标函数基于图像和图像的地面实况标签序列来获取成本值与之相对照在其上进行端到端的训练。

在本次项目的训练日志中，也可以明显地发现利用随机梯度下降的训练方法的特征。

通过 ADADELTA 优化相比传统的动量方法由更快的收敛速度，实现更好的效果。通过反向传播算法来计算其所有的梯度特征，尤其是在转录层中，通过误差微分用前向后向的算法反向传播。在循环层中，误差微分是通过时间反向传播的数据来计算获得的。通过 ADADELTA 自动计算每维的学习率从而达到优化目的。

2.2.6 CRNN 模型训练数据集

卷积递归神经网络模型使用了开发者们发布的合成数据集作为训练数据。该数据集包含八百万张训练图像及其对应的场景实况词。训练数据进行任何微调的操作成为了可以忽略的行为，网络便可以利用这些合成数据进行训练，并在所有其他真实场景的测试数据集上进行了测试。由于数据集图像由合成文本引擎生成，并且非常逼真，因此在实际场景下得到的结果也是十分优越。

2.2.7 CRNN 模型总结

卷积递归神经网络融合了卷积神经网络和循环神经网络的优质特性。相比较于传统方法以及其他基于卷积神经网络和循环神经网络的算法，在场景文本识别基准的实验中卷积递归神经网络实现了极为优秀的性能。不仅如此，卷积递归神经网络在光学音乐识别的基准上也明显优于其他竞争对手，这验证了卷积递归神经网络的普遍性，证实了所提出算法的杰出表现。这些优越的特性令卷积循环网络成为基于图像的序列识别的极好方法。CRNN 模型在来自标准文本识别基准的真实图像上效果很好，即使其是完全使用合成文本数据进行训练。

卷积递归神经网络可以应用于各种实际场景，也就是说它也可以应用于例如古代文字辨义等其他领域的问题，也包括图像中的序列预测这类情况。使卷积递归神经网络在场景应用中更加实用，在满足目标准确率要求的情况下一步加快卷积递归神经网络的速度是未来值得探索的方向之一。

2.3 NLP 自然语言处理

通过 OCR 处理模型成功实现了文本检测与文本识别功能，并能输出文本类型结果用于后续处理。在接下来的设计目标报告评分中，也就是在自然语言处理的机器学习大类范围中，便可以合理的将本次项目实现功能具体化为实现自然语言评估功能。

在过去的几年里，自然语言生成(NLG)的研究取得了巨大的突破，模型现在能够翻译文本、总结文章、从事使用越来越复杂的方法，以极其夸张的准确度对图片进行对话和讨论甚至辩论。目前有两种方法可以评估这些 NLG 系统：人工评估和自动度量。通过人工评估，人们会使用人工方法的标签化处理对模型的每个新版本进行大规模质量调查，但这种方法可能会耗费大量人力。相比之下，可以使用流行的自动度量(例如，BLEU)，但这些通常是人类解释和判断的不可靠替代品。NLG 的快速发展和现有评估方法的缺陷要求开发新的方法来评估 NLG 系统的质量和成功率。

NLG 的评估系统分为自动评估和人工评估。在人工评估中，将一段生成的文本呈现给注释者，他们的任务是评估其流畅度和含义方面的质量。文本通常与参考并排显示，由人类创作或从 Web 挖掘。这种方法的优点是准确，因为人们在评估一段文本的质量仍然是无与伦比的。但是，这种评估方法很容易花费数天时间，对于涉及几千个示例的样本就要求多个人进行操作，而这会破坏模型开发工作流程。

自动指标的优势和劣势与人类评估的优势和劣势相反。自动指标很方便，它们可以在整个训练过程中实时计算。然而，它们往往是不准确的，因为它们专注于表面的相似性，而且它们不能捕捉人类语言的多样性。通常，有许多完全有效的句子可以表达相同的意思。即使面临下面说的这种指标没有准确地捕捉到意义，而非正确地对其他转述的情况下，完全依赖词汇匹配的重叠指标这种度量标准也是无法正确地评估那些在表面形式上与参考文献相似的句子。

因此，理想情况下，NLG 的评估方法应该结合人工评估和自动度量的优势——它的

计算成本应该相对较低，但同时也具备足够灵活的特性以应对语言多样性。

根据^[12]中内容可以看出，在过去的几年里，基于自然语言的文本生成已经取得了重大进展。然而评价方法已经落后了，因为最流行的选择（例如 BLEU 和 ROUGE）可能与人类的判断相关性很差。开发者提出了 BLEURT，一种基于 BERT 的学习型评价方法，它可以用数万个可能是双关的训练实例来较为准确地模拟人类的判断。关键的改良是使用一个全新的预训练方案，它通过数百万个合成的例子的数据协助来实现模型泛化。与普通的基于 BERT 的方法相比，BLEURT 即使在训练数据稀少和分布不均的情况下，它也能产生卓越的再结果。

目前的方法主要分为两类完全学习的度量方法，如 BEER、RUSE 和 ESIM 是端到端的训练，它们通常依赖于手工制作的特征和/或学习的嵌入。另一种截然不同的方法是混合度量，如 YiSi 和 BERTscore，将训练过的元素，如上下文嵌入，与手写逻辑相结合，如标记对齐规则。第一类通常提供了很大的表现力，如果有一个人类收视率数据的训练集，度量标准可以充分利用它，并紧密地适应收视率分布。此外，学习到的度量可以被调整为测量特定任务的属性，如流畅性、忠实度、语法或风格。另一种方法混合度量提供了稳健性。在几乎没有训练数据的情况下，它们可以提供更好的结果，并且他们并不依赖于培训和测试数据是一致分布的设定情况。

事实上，在 NLG 评估中存在一些特别的问题，因为领域漂移是衡量标准文献的主要目标，但在这种情况下也可能产生质量漂移。NLG 系统往往会随着时间的推移而变得更好。因此，在 2017 年的评级数据上训练出来的模型可能无法区分 2022 年的顶级系统，尤其是对于较新的研究任务。一个理想的学习型指标既能充分利用现有的评级数据进行训练，又能对分布漂移保持稳健，也就是说，它应该能够外推。我们的见解是，通过在大量的合成数据上预训练一个完全学习的指标，然后再在人类的评分上进行微调，就有可能把进步性和鲁棒性结合起来。为此目的，引入了 Bleurt 作为一个文本生成指标。一个关键的因素是 Bleurt 特征是一个新颖的预训练方案，它使用维基百科句子的随机扰动，并辅以一组不同的词汇和语义层面的监督信号。测试表明，它可以很容易地适应不同的工作，从数据到文本的数据集的三个任务。实验表明，合成预训练方案在 IID 环境中提高了性能，并且在训练数据稀少、倾斜或域外时对确保鲁棒性至关重要。

3 文本检测与识别的具体实现过程

3.1 对于现有模型的应用测试

3.1.1 本次项目中所试用过的文本检测方法

Opencv: 语言框架和中文检测效果不够理想

CascadeTabNet: 对于中文文字字符的检测效果较差

image2csv: 对于数字字符识别效果较好，而对于文字文本的检测识别效果较差

3.1.2 本次项目中所试用过的文本识别方法

Opencv: 语言框架和中文检测效果不够理想

FeatureCamp 文本识别：难以进行模块集成，识别效果也不够理想

paddleOCR：初始模型对于手写字符的识别效果很差

3.2 基于 Paddle 开发者工具实现手写报告文本检测与识别功能

经过多次对比测试后在 tensorflow 和 paddle 中因为中文字符的识别匹配度和可靠性选择通过 paddle 来实现本次手写报告文本检测和识别功能。

3.2.1 文本检测

通过 paddleOCR 开发者工具库分别测试调试了文本检测算法 DB、EAST、SAST 三种算法，同时结合设备性能分别调试了以下文本检测模型：

```
ch_PP-OCRv2_det_slim
ch_PP-OCRv2_det
ch_ppocr_mobile_slim_v2.0_det
ch_ppocr_mobile_v2.0_det
ch_ppocr_server_v2.0_det
```

发现 ch_PP-OCRv2_det_infer 模型对于所测试的文本报告手写字体检测内容效果更好，能够较为准确以低错误率的方式正确识别出报告版式以及手写文本内容，满足本次项目所需条件。

3.2.2 文本识别

基于 paddleOCR 开发者工具库

结合设备性能测试调试了文本识别算法：

```
ch_PP-OCRv2_rec_slim
ch_PP-OCRv2_det
ch_ppocr_mobile_slim_v2.0_det
ch_ppocr_mobile_v2.0_det
ch_ppocr_server_v2.0_det
```

经过测试发现以上模型对于所测试的文本报告字体识别效果较差，通过结合已有设备条件，准备通过训练新的模型来提高对于手写实验报告的识别准确率以优化识别效果。

3.3 文字检测与识别深度学习具体训练过程

3.3.1 数据集搜集与标签化整理

为了提升所适用模型的实用性和可靠性，同时为了项目内容的可复现性和可验证性，舍弃完全依靠个人数据集的策略，通过采取公共数据集的方式来获取可靠的数据集内容。

常用的手写文字公共数据集有中科院自动化研究所手写中文数据集和 NIST 手写单字数据集。

综合考虑样本量，数据集数据来源，样本字符条件等多种因素，经过评估最终选择了使用中科院自动化研究所-手写中文数据集作为本次文本识别模型的训练用数据集。

中科院自动化研究所手写中文数据集 CASIA Online and Offline Chinese Handwriting

Databases 集成的复杂数据内容所包含的数据信息并不全是本次训练中所必要的，本次模型训练目标需要部分数据信息。为此通过 python 程序将所获取的数据集进行标准化处理，从中分离得到图片文件与标签文件并整理为 Simple Dataset 数据格式，用于训练。通过 GTLC tool 标签工具可以自行添加并扩大数据集内容，在本次项目中由于时间和完整性考虑选择直接应用分离出的图片与标签文件作为数据集数据。

3.3.2 预训练模型

由于训练模型是基于预训练模型在真实数据与竖排合成文本数据上 finetune 得到的模型，在真实应用场景中有着更好的表现，预训练模型则是直接基于大量真实数据与合成数据训练得到，更适合用于在自己的数据集上 finetune，因此选择通过引入预训练模型的方法节约训练模型所需要的时间，同时提高模型对于中文文本的识别敏感性。

通过对比测试和验证性能的对比，选择适配训练的模型 ch_ppocr_mobile_v2.0_rec_train 作为预训练模型使用。

3.3.3 训练模型网络结构设计

具体训练实现流程如图 3-1 所示。

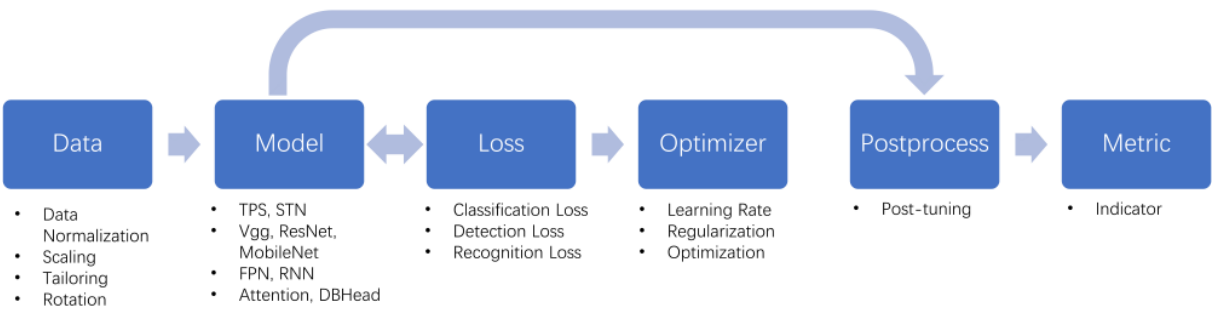


图 3-1 训练实现流程

实际网络构建如图 3-2 所示。

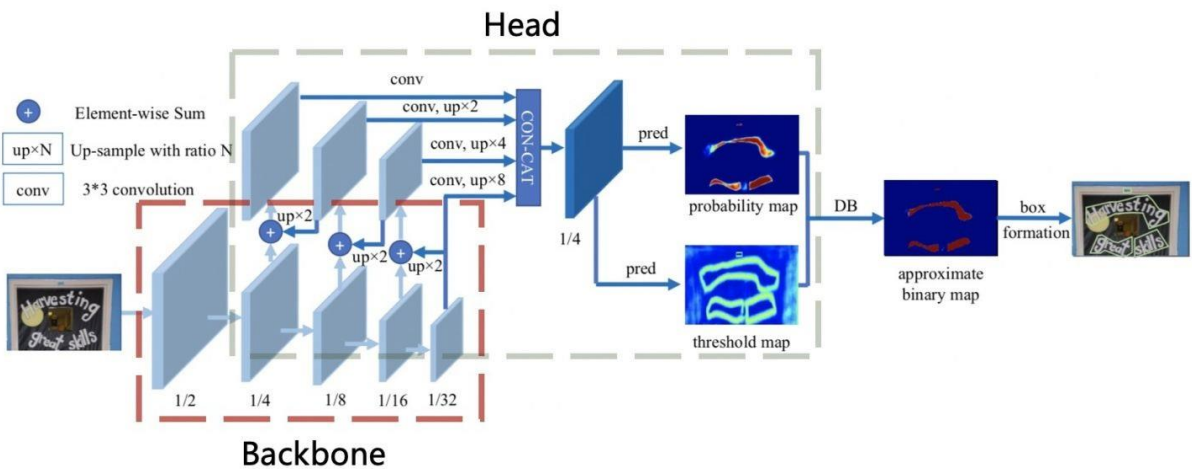


图 3-2 训练网络构建图

在本次训练网络中，网络被划分为 Transform, Backbone, Neck 和 Head 四个阶段。

Backbone 部分由 MobileNetV3 组织构成，Neck 部分由 SequenceEncoder 构成基于 CRNN 算法，Head 部分由 CTCHead 构成。

通过多次训练调参，将训练中的 batchsize 单卡设置为 256，由此能更好地适配 GPU 的计算逻辑。

训练迭代相关主要参数设置如下：每个训练数据载入装载 936 iters；每 10 个 iteration 记录一次 loss，准确率等参数；每次 epoch 保存一次训练输出的模型和对应数据；一共进行 100 次 epoch，能有效地提升训练的模型性能；每 2000 次迭代评估一次。

3.3.4 训练结果与过程记录

经过训练最终生成模型，记录下的日志文件损失和准确度结果如图 3-3,3-4,3-5 所示。

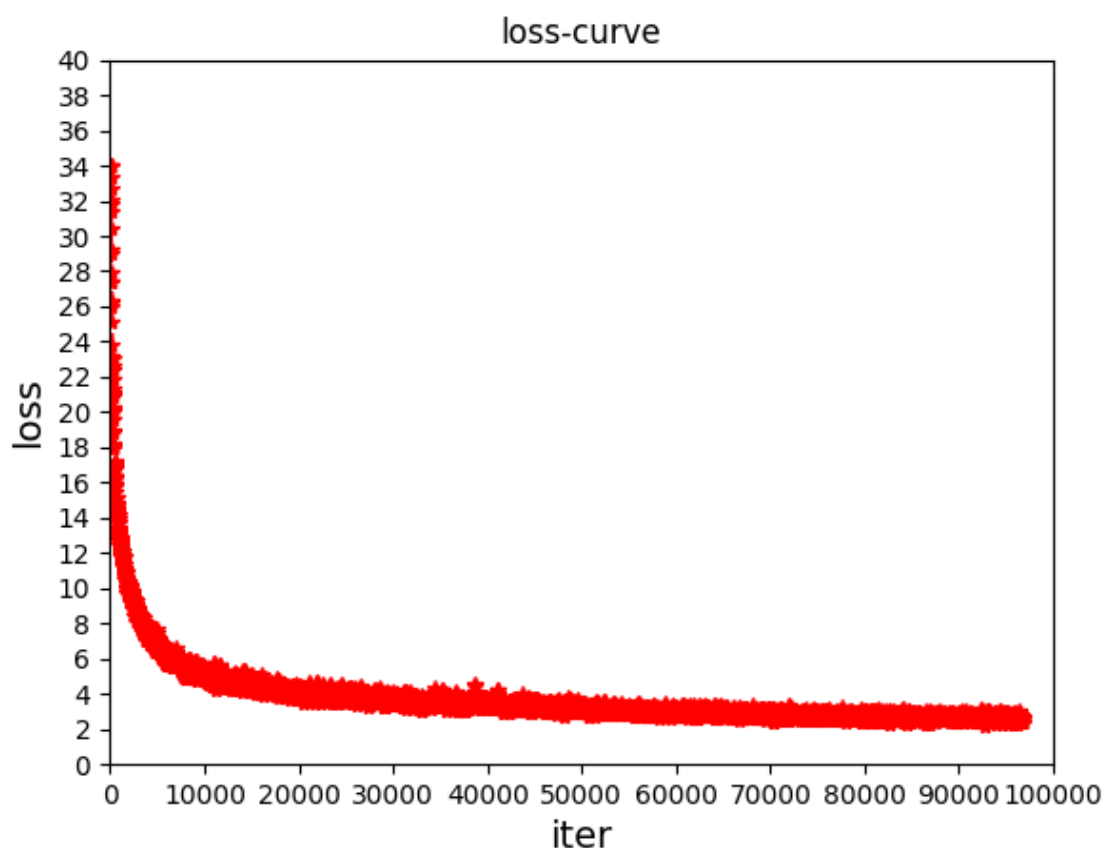


图 3-3 训练损失随迭代次数变化曲线图

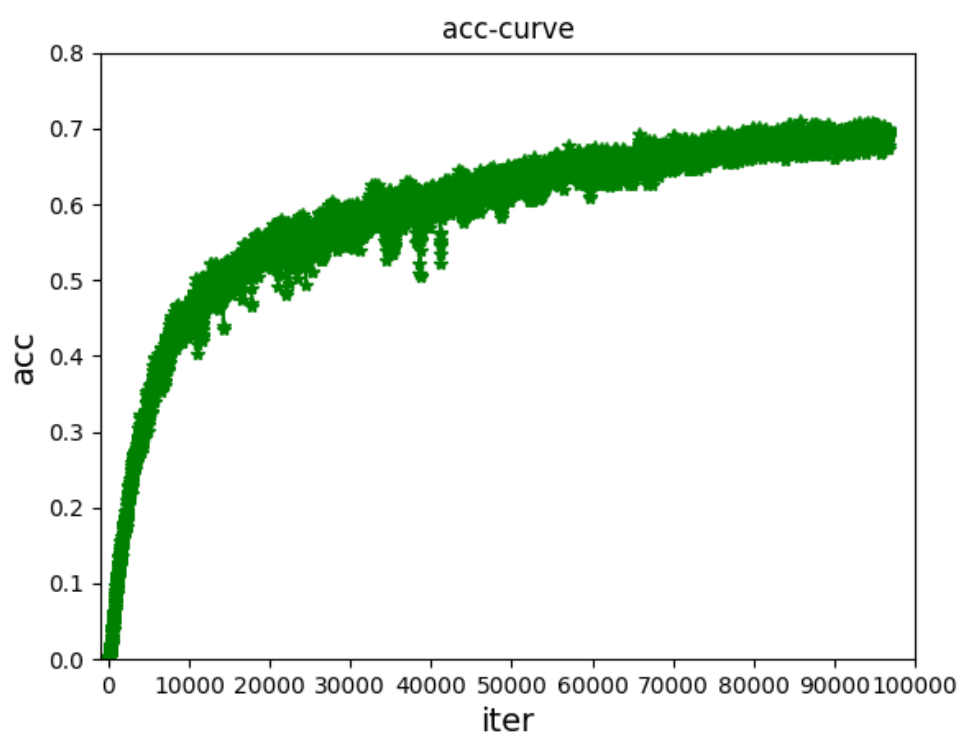


图 3-4 模型准确率随迭代次数变化曲线图

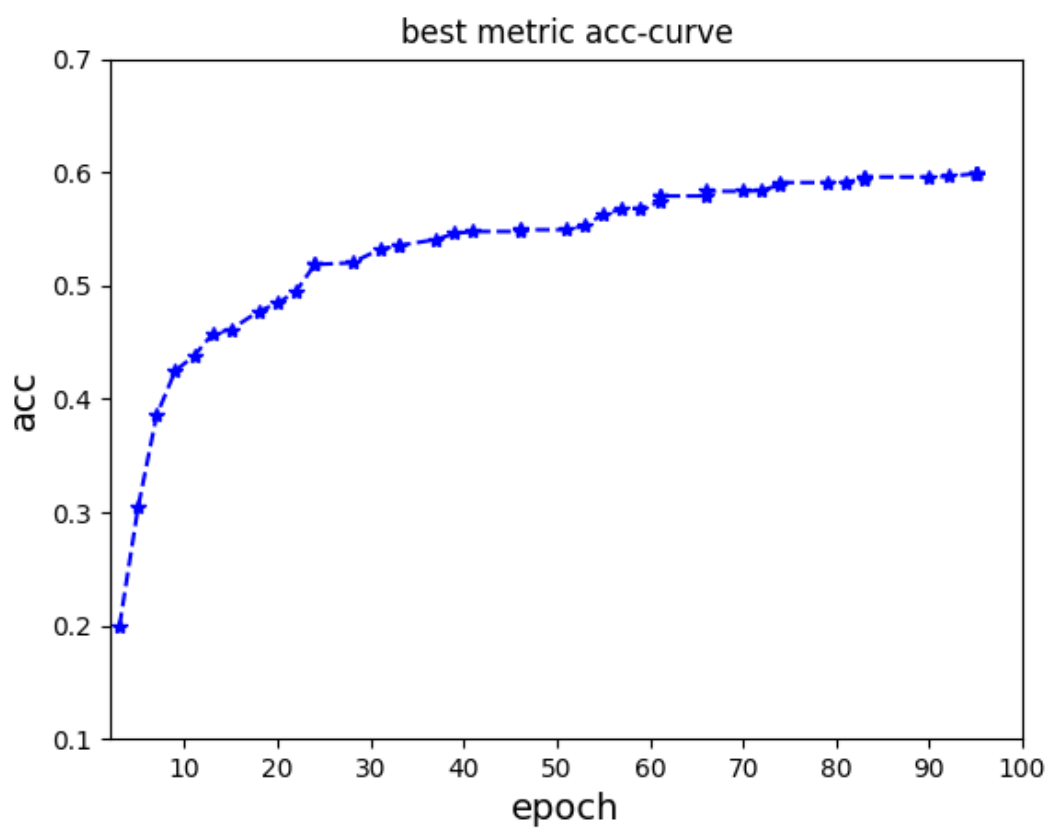
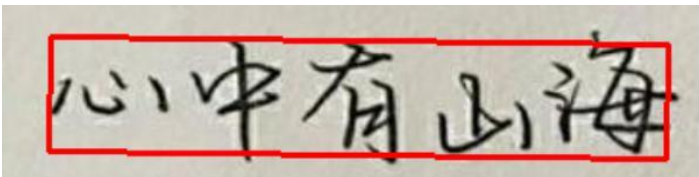


图 3-5 训练最佳模型准确率图表

从图 3-3 中可以看出经过迭代，模型的 loss 在 1 万次迭代过程中有一个明显的梯度下降，经过 10 万次迭代训练后基本稳定于 2 的低值。同时 loss 图表所显示的损失梯度和下降情况也对应了所设计的随机梯度下降的方法，验证了此方法的合理性与有效性。从图 3-4 可以看出模型在训练中的几个偏移点，同时看出模型经过训练后最终准确率在提升至 60%左右后基本保持在稳定水平，图 3-5 所示最高准确率模型曲线大致与全过程的模型训练准确率变化状态相近，说明收敛效果明显，由此可以判断训练效果很好。在此过程中两个图像对比并没有明显幅度大的差异部分，说明模型的演化方向准确，没有发生雪崩式偏移情况。为了对模型实际训练后效果进行验证，将训练所得模型与初始轻量模型和预训练模型进行对比测试并记录结果。

对于单行文本的识别效果如表 1 所示。

表 1 单行文本识别结果

模型列表	识别文本结果
初始轻量模型	 1: 心有的 0.760
预训练中文识别模型	 1: 心1中有山海 0.685
训练后中文识别模型	 1: 心中有山海 0.998

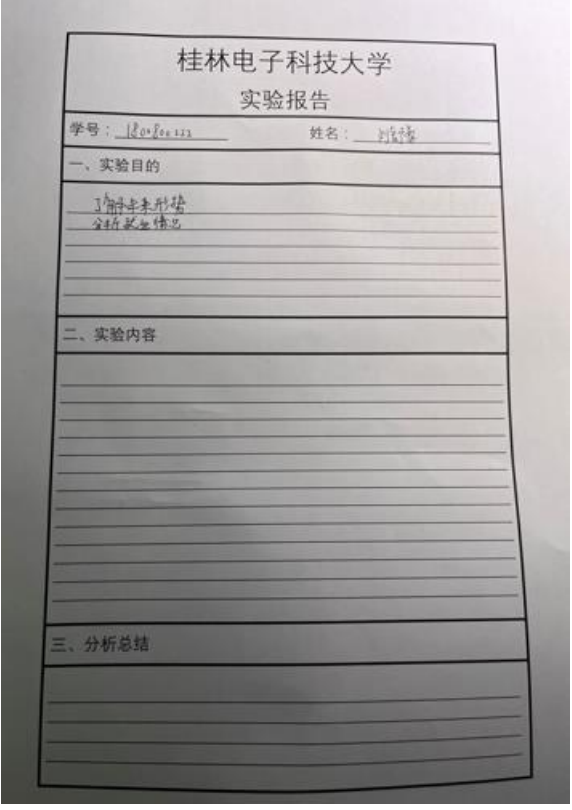
对于清晰手写图片检测识别效果如表 2 所示

表 2 清晰图片检测识别结果

识别图片	模型	识别结果及参考准确率
	初始轻量模型	1.牧林电子科技大学 (0.848) 2.实验报告 (0.951) 3.学写 (0.837) 4.1800800 工 2 (0.730) 5.姓名 (0.893) 6.实验目的 (0.887) 7.掌握单片机起德团 (0.612) 8.实验内容 (0.933) 9.分析总结 (0.996)
	预训练中文识别模型	1.桂林电子科技大学 (0.999) 2.实验报告 (0.999) 3.学号 (0.996) 4.1800800222 (0.976) 5.姓名 (0.819) 6.刘子豪 (0.942) 7.实验目的 (0.998) 8.学握单片机使用 (0.718) 9.实验内容 (0.997) 10.训练模型 (0.872) 11.分析总结 (0.918)
	训练后中文识别模型	1.桂林电子科技大学 (1.000) 2.实验报告 (1.000) 3.学号 (1.00) 4.1800800222 (0.988) 5.姓名 (1.000) 6.刘舒豪 (0.889) 7.实验目的 (1.000) 8.掌握单片机使用 (0.919) 9.实验内容 (1.00) 10.训练模型 (0.997) 11.分析总结 (1.000)

对于潦草书写图片检测识别效果表 3 所示。

表 3 潦草书写图片检测识别结果

识别图片	模型	识别结果及参考准确率
	初始轻量	1.牧林电子科技大学 (0.917) 2.实验报告 (0.946) 3.学告。1800800222 (0.745) 4.姓名“到短” (0.720) 6.实验目的 (0.897) 7.了解至半来形势力 (0.791) 8.实验内容 (0.889) 9.分析总结 (0.993)
	预训练模型	1.桂林电子科技大学 (0.999) 2.实验报告 (0.999) 3.学号：1800800222 (0.858) 4.姓名：刘子 (0.860) 5.实验目的 (0.999) 6.1 角子来形垫 (0.514) 7. 分就也情区 (0.630) 8.实验内容 (0.863) 9.分析总结 (0.996)
	训练后模型	1.桂林电子科技大学 (1.000) 2.实验报告 (1.000) 3.学号：1800800222 (0.978) 4.姓名：刘舒象 (0.831) 5.实验目的 (1.000) 6. 用手分来形垫 (0.567) 7.分主斤就也情况 (0.826) 8.实验内容 (0.999) 9.分析总结 (1.000)

将所识别到的文本内容按照行存放在 txt 文件中，通过识别关键字的方法将实验报告内容和标题等印刷的关键字区分，根据报告内容使用者需求可以自行进行设定处理。这个过程目的是为了格式化处理经过检测和识别后的报告内容，便于应用的自然语言处理模块进行评分处理。

4 基于 BLEURT 的 NLP 文本评分

本次所应用的 BLEURT 评估方法是一种新颖的、基于机器学习的自动度量，可以捕获句子之间的重要语义相似性。它在公共评级集合（WMT 指标共享任务数据集）以及用户提供的其他评级上进行训练。Bleurt 模型是一个自然语言综合组织的评估模型。他通过导入一组对应的语句，一个对照文件和一个基准文件，他返回一个分数来显示出引用文件的文本流畅和正确水平，同时表示出这个引用文件表达出的基准文件的内容意义的程度。Bleurt 是一种经过训练的评估模型，也就是说，它是一种通过衰退模型，在已评估的数据集上基于 BERT 和 RemBERT 训练而得到的评估手段。

在单句的对比性能上此模型可以很好地进行表意语句的相似性对照，具有极高的可靠性。通过设置参考文本为“掌握使用电源模块的方法”进行测试，结果如表 4 所示。

表 4 评估文本及其结果

序号	评估文本	评估结果
1	掌握使用电源模块的方法	INFO:tensorflow:BLEURT initialized. [1.0012092590332031]
2	掌握使用单片机的方法	INFO:tensorflow:BLEURT initialized. [0.5198370814323425]
3	复习模电知识	INFO:tensorflow:BLEURT initialized. [0.3324030041694641]
4	学会使用电源模块的手段	INFO:tensorflow:BLEURT initialized. [0.7354535460472107]

由 1 看出模型很好的识别出了两个相同文本的效果给出了满分。由 2 看出评分明显下降，说明模型正确辨识出了样本结果与参考答案的区别。当句意进一步偏离时。由 3 看出评分结果分数进一步下降。当表述方式改变而句意相同时，测试模型能否正确识别出两者的语义关联性。由 4 看出即使相同的字符并没有重复出现，此句测试结果仍然高于 2 的结果，说明模型具有良好的语言识别和语境处理能力。模型也可以同时进行多句多段的综合评估，通过进行多个句子情况，以及长句报告表述情况下的对比检验评分，测试判断模型对此类样本的分析灵敏度。参考文本为“掌握使用电源模块的方法，并能自主进行通电测试。”评估文本设置为“学会使用电源模块的手段，同时能自己进行通电测试。”得到如图 4-1 所示结果。

INFO:tensorflow:BLEURT initialized.
[0.7701483368873596]

图 4-1 多句长句文本评估结果

从表 4 结果可以看出，模型对于多句长句的处理能力仍然优秀。在实际应用场景中，需要进行全文的报告评估。结合系统前面所获取的文本检测与识别模块的输出内容，测试模型是否能很好地处理多行多段语句。不妨设检测识别输出文本如图 4-2 所示。

```
OCRresult.txt X
1 桂林电子科技大学实验报告
2 学号
3 1800800222
4 姓名
5 刘舒豪
6 实验目的
7 掌握单片机使用。
8 实验内容
9 训练模型
10 分析总结
11
```

图 4-2 文本检测识别输出结果

当参考文本与其相同时，评估结果如图 4-3 所示。

```
INFO:tensorflow:Loading model.
INFO:tensorflow:BLEURT initialized.
[0.9597343802452087,
0.8378193974494934,
0.9673959612846375,
0.8770206570625305,
0.9382616877555847,
0.952293872833252,
0.9963065981864929,
0.968858003616333,
0.9563018083572388,
0.9370981454849243]
```

图 4-3 同文本对比检测结果

从以上模型评估结果可以看出当评估文本和参考文本全都一致时，评估均得到了 85 分以上的高分值，通过更改参考文本使得参考文本与评估文本一处有差异再进一步测试。

```
Answerresult.txt X
1 桂林电子科技大学实验报告
2 学号
3 1800800222
4 姓名
5 刘舒豪
6 实验目的
7 掌握单片机使用，学会使用电源模块的手段。
8 实验内容
9 训练模型
10 分析总结
11

INFO:tensorflow:BLEURT initialized.
[0.9597343802452087,
0.8378193974494934,
0.9673959612846375,
0.8770206570625305,
0.9382616877555847,
0.952293872833252,
0.49583059549331665,
0.968858003616333,
0.9563018083572388,
0.9370981454849243]
```

图 4-4 差异参考文本

由图 4-4 文本内容的对比检测结果可以看出模型很好区分出了全文中语义不同的地方，并且准确地判断出其语义缺失的问题，给出了相对较低的评分。

综上，可以判断出模型具有良好的单句，长句，多句，以及全文多行多段的自然语言处理和语义判断能力，达到了预期的目标，成功地实现了对报告进行评分的功能。同时可以通过个性化整合的方式，结合文本检测与识别的输出设定方法，将所设定的“实验目的”等版面标志词汇与手写报告答案内容进行区分处理，将这些分值根据个人需要进行加权平均，便可以得到理想的手写报告评分结果。

5 基于服务器实现的网页形式前端客户端设计

用户交互系统流程设计如图 5-1 所示。

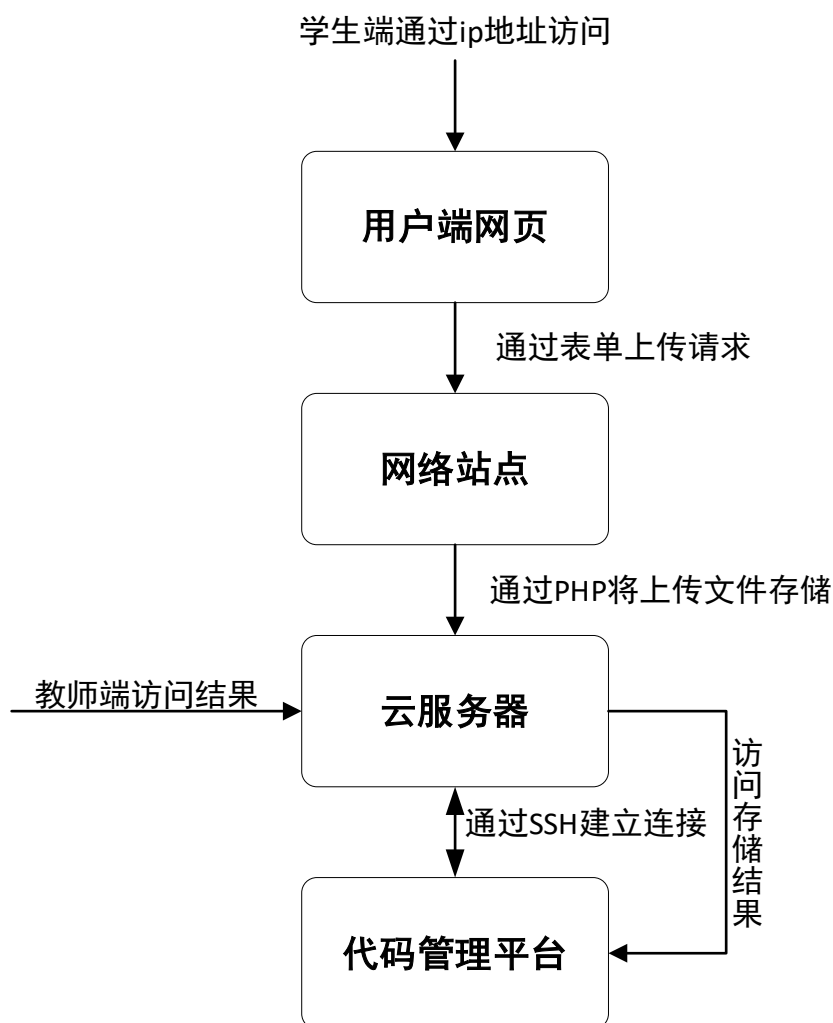


图 5-1 用户交互系统整体流程图

通过 Dreamweaver 对前端网页界面与功能进行设计编辑，主要由网页主页部分和文件操作部分构成。

用户通过基于表单的提交方法将需要进行评估的报告文件进行提交，同时，利用 js 脚本获取所提交文件的路径并提供预览效果。提交后的表单经过转发，通过 php 的方法将其存放在服务器指定目录下，服务器的管理者可以自行处理这些报告图片或者将其打包转发给教师用户。

在本次功能设计中，实现了通过服务器利用 SSH 连接的方法访问 Google Colab 的功能，而文字识别处理与自然语言处理模块部署在 Google Colab 在线代码保管平台。通过平台化管理的方法便于在多种环境和实际应用场景下进行手机，电脑等多种终端的调试、运行、模型训练等操作。

JS 全称为 JavaScript,是一种由 Netscape 的 LiveScript 发展而来的脚本语言。这种语言要目的是为了解决服务器终端语言遗留的速度问题。在本次项目中为了使得上传的结

果同步显示在页面中，主要通过“onchange”语句在用户改变 input 输入框的内容时执行一段 JavaScript 代码，便于提交报告的学生或者其他用户对上传文件做进一步的确认，以减少使用者上传错误文件或者上传文件损坏的情况发生。

网页主页设计界面如图 5-2 所示。

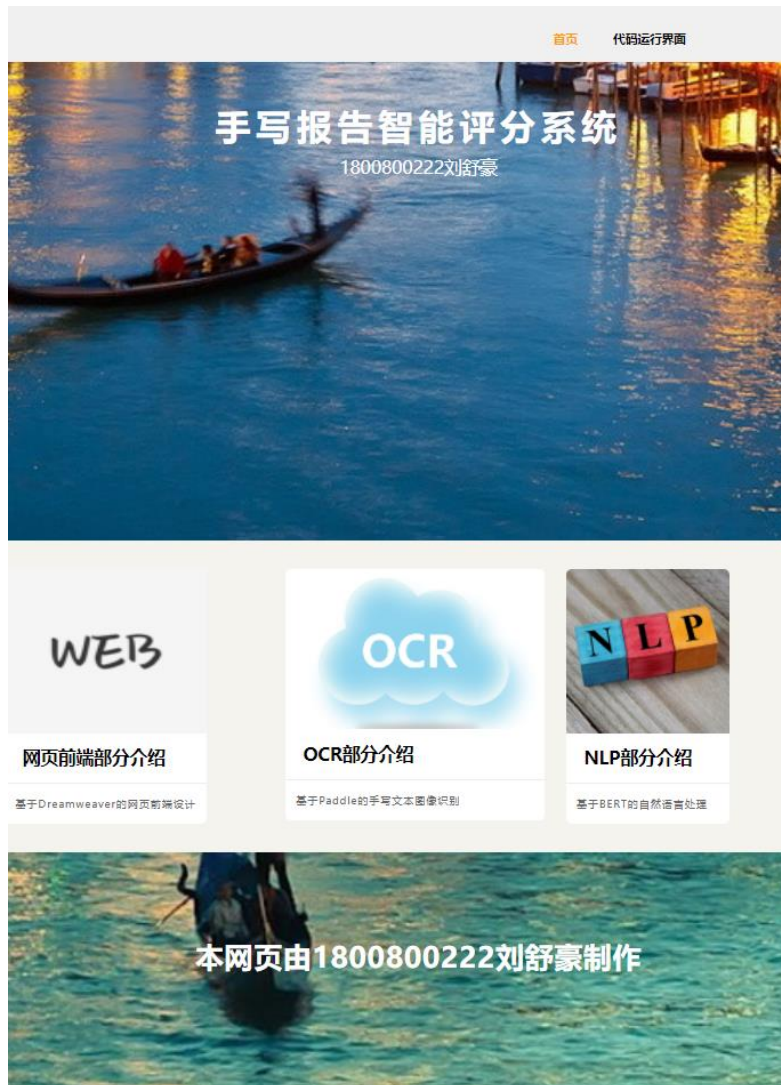


图 5-2 网页设计主页界面

代码执行界面如图 5-3 所示。

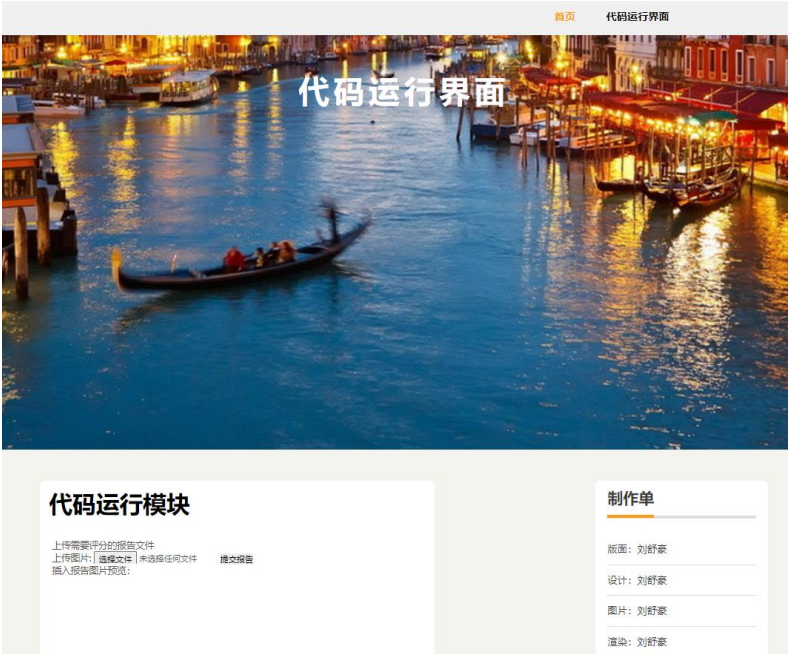


图 5-3 图片上传与预览页面

界面上上传图片成功实现预览效果如图 5-2 所示。

代码运行模块

上传需要评分的报告文件
上传图片 | 选择文件 | baogao666.png 提交报告
插入报告图片预览:

桂林电子科技大学 实验报告	
学号	1800800222
姓名	刘野豪
实验目的	学习单片机使用 单片机引脚输入输出方法 和接口原理
实验内容	掌握软件操作办法 测试记录程序运行结果
分析总结	懂得如何使用硬件

图 5-4 网页上传预览效果

提交文件服务器响应结果如图 5-5 所示。

```
458722上传文件名: baogao666.png
文件类型: image/png
文件大小: 447.970703125 kB
文件临时存储的位置: /tmp/phpgdFzH2
文件存储在: filebox/baogao666.png
```

图 5-5 服务器提交文件响应效果

服务器上传文件成功结果如图 5-6 所示。

根目录 > www > wwwroot > fileget.com > filebox >

上传

远程下载

新建

文件查找

收藏夹

分享列表

终端

/(根目录) (32G)

<input type="checkbox"/>	文件名	权限 / 所有者	大小	修改时间	备注
<input type="checkbox"/>	DrinkOrder.txt	644 / www	122 B	2022/04/20 17:26:05	
<input type="checkbox"/>	baogao666.png	644 / www	447.97 KB	2022/05/13 12:14:03	

图 5-6 服务器端文件列表

SSH 又称为安全外壳协议，由 IETF 的网络小组（NetworkWorkingGroup）所制定。对于客户端来说，主要有两种级别的安全认证方式，基于口令认证的第一级别和基于密钥的第二级别。而 SSH 这项安全协议建立在应用层基础上，同时是一种具有高可靠性，高保密性，确保远程登录会话等服务安全进行的协议。在本项目服务器与代码管理运行平台 GoogleColab 的连接中，具体是通过基于密钥的 SSH 公钥方式实现的。在本项目中这种认证手段因其具有的便捷性和可靠性，可以有效地从服务器端获得 GoogleColab 端的有效权限，便于调用代码运行或者进行模型优化、训练等后续处理等优点，经过抉择后应用于服务器与 Colab 的验证对接中。

服务器连接成功后访问效果及代码服务器文件列表如图 5-7 所示。

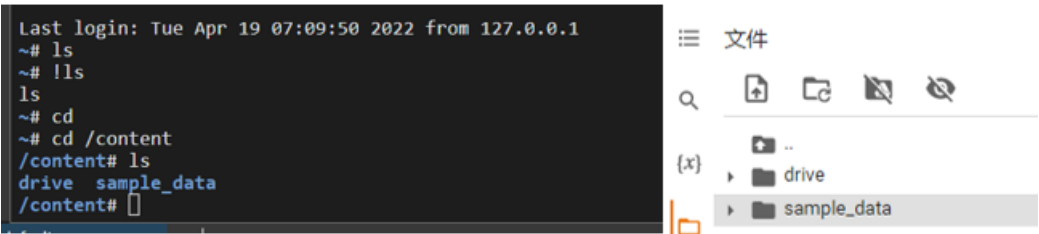


图 5-7 服务器访问效果

经 Google Colab 模型验证图片结果如图 5-8 所示。

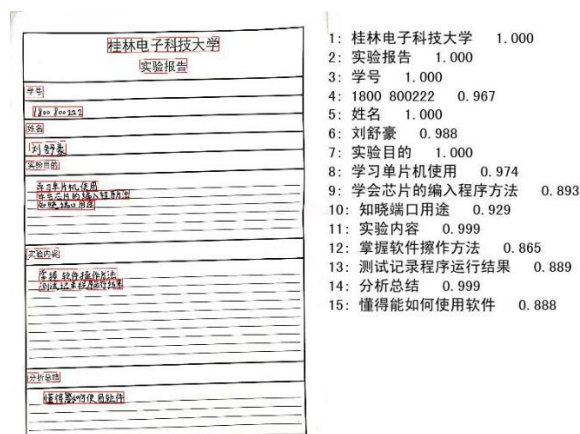


图 5-8 上传图片检测识别结果可视化

综上，系统顺利实现了从客户端到网站服务器到代码管理平台的一体化管理，提供了个性化定制实验报告格式的手段，为教师端设置实验报告标准答案提供便利，只需将标准实验报告答案按照所设置的格式存放在系统程序所指向路径的 txt 文件中即可。从测试结果看出模型具有良好的稳定和可靠性，文本检测识别与报告评分功能均达到预期。

6 结论

随着课堂教学中实验变得愈发重要，实验在教学中的评估比重和学分时长进一步增加，因此教师实验报告的批改的时间花费和错误发生率也同步提升。由于人工智能的进一步发展，其在文本检测识别处理和自然语言处理的领域内发展的技术提供了对于手写实验报告批改方式的新途径。然而，现有的文本检测识别处理模型和自然语言评估模型并没有很好地实现集成化效果，从而无法系统性地处理此类问题。

本文通过深度学习的方法进一步提高模型对于混用文本的检测识别能力，并将手写字符文本检测识别与报告评分应用场景结合在一起，成功实现并测试了以下内容：

- (1) 评估并测试了多种文本检测和文本识别的方法，通过组合的方式利用性能最佳的模型实现对于手写报告文字的识别与格式化输出。
- (2) 解决了进行识别模型训练，测试并调整参数基于 paddle 工具通过 CRNN 成功训练了对于手写字符文本识别率较高的文本识别处理模型。
- (3) 解决了对于获取的报告文本内容进行评分的问题，调试并测试了基于 Bert 预训练的语言表征模型的 Bleurt 自然语言处理评分模型。
- (4) 解决了系统与学生、教师之间的实验报告交互问题，设计并实现了基于 html 和 js 的图片上传、预览的前端网页页面，通过 php 设计并实现了服务器端接收和存储的功能，通过 ssh 验证连接实现了服务器端与代码保存管理工具的对接功能。

本次系统设计仍存在可以发展改良的空间，例如在深度学习中训练的模型数据集样本量相对而言不够充足，可以通过搜集更多实验报告样本文件进行训练从而提升模型识别效率与效果。模型对于多种语言均有识别功能，但本次任务中只对于中文文本进行了训练和调试测试，经过扩展训练可以实现对多种语言的支持效果等。

谢 辞

在毕业设计的过程中，遇到了新接触的知识，同时也学习了许多新的编程方法，网站搭建设计方法和专业软件的使用手段。在近及毕业之时，我对于在本次设计项目中那些帮助过我的老师，同学，以及在各方面提供支持的学校表示衷心的感谢。

首先要感谢我的毕业设计的指导老师首照宇老师。本次课题的理念设计和内容构成都是由首照宇老师精心设计的，同时首照宇也根据本科生的技术水平进行合理的任务分配和时间计划，使得我的毕业设计项目流程清晰，推进平稳。

其次要感谢文辉老师在我的毕业设计中对于我的各种疑问和困惑都知无不言地耐心回答，同时在毕业设计的转折点和方案选择上提供了建设性的参考意见和方向指导，使得我的毕业设计进程顺利并且十分有效率，在此过程中也指导了许多相关软件和操作方法。

同时也要感谢各位陈怡欣和谢明权等学长在毕业设计过程中对我的帮助、鼓励和支持，使得我在遇到课题困难时能够获得指导和解惑。

参考文献

- [1] 王庭栋.基于稠密卷积网络的文字识别研究[J].科学技术创新,2021(20):89-90.
- [2] Baoguang Shi,Xiang Bai,Cong Yao. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition.[J]. CoRR,2015,abs/1507.05717.
- [3] 陆金江.基于数字图像处理的字迹识别技术探讨[J].兰州工业学院学报,2020,27(05):52-55.
- [4] 徐富勇,余谅,盛钟松.基于深度学习的任意形状场景文字识别[J].四川大学学报(自然科学版),2020,57(02):255-263.
- [5] 孙峰.深度学习视角下关于合理利用教学资源的探究[J].大庆社会科学,2022(01):147-150.
- [6] 李新炜,殷韶坤.深度学习在文字识别领域的应用[J].电子技术与软件工程,2018(24):40.
- [7] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.
- [8] 夏昌新,莫浩泓,王成鑫,王瑶,闫仕宇.基于深度学习的图像文字识别技术研究与应用[J].软件导刊,2020,19(02):127-131.
- [9] 任逍航.基于深度学习的自然场景图像的中文字检测 and 识别算法研究[D].上海交通大学,2017.
- [10] 卢未来.面向图像的场景文字识别技术研究[D].辽宁工业大学,2018.
- [11] 高威威.基于深度学习的自然场景文字识别[D].安徽大学,2019.
- [12] 罗泉.基于深度学习的自然语言处理研究综述[J].智能计算机与应用,2020,10(04):133-137.
- [13] 蒋萍.基于深度学习方面自然语言处理技术(NLP)的研究[J].数字通信世界,2021(01):31-33.
- [14] 江洋洋,金伯,张宝昌.深度学习在自然语言处理领域的研究进展[J].计算机工程与应用,2021,57(22):1-14.
- [15] Yoav Goldberg,车万翔,郭江,张伟男,刘铭.基于深度学习的自然语言处理[J].中文信息学报,2021,35(08):145.
- [16] 马迅飞,张进铎,张卫华,张红杰,杨显峰.基于 SSH 框架技术的 IT 运维管理系统研究与开发[J].信息技术与信息化,2020(06):51-53.
- [17] 阎适成,孟艳红,王泽京.基于 SSH 框架的爱心捐赠系统的设计与实现[J].智能计算机与应用,2020,10(06):98-100.
- [18] 陆金江.基于深度卷积神经网络的模糊字迹图像识别方法[J].佳木斯大学学报(自然科学版),2021,39(05):42-46.
- [19] Thibault Sellam. Dipanjan Das. Ankur P Parikh. BLEURT: Learning Robust Metrics for Text Generation. Proceedings of ACL. 2020
- [20] Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Trans. Pattern Anal. Mach. Intell. vol 39. no 11. pp. 2298-2304. 2017