

**计算机科学技术学院**

**本科生毕业论文**

**基于OCR的身份证要素提取设计与实现**

**班 级： 云计算数据中心运营**

**学 号： 20161104492**

**姓 名： 高尚**

**指导教师： 李艳玲**

**2020 年 3 月 1 日**

摘要 本论文针对OCR系统中身份证复印件设计一套身份证要素信息提取系统，系统采用前后端分离技术，前端系统负责上传图片和展示识别结果，后端系统负责图像处理和文字识别；提取系统具备低耦合特点，各个子模块可以独立运行，文字识别模块也可用于其它文本识别任务。

关键字 OCR、深度卷积神经网络、图像识别

目录

**[1 绪论 4](#_Toc18491)**

[1.1 研究背景 4](#_Toc7484)

[1.2 研究难点 5](#_Toc26158)

[1.3 研究内容 5](#_Toc6949)

**[2 系统流程介绍 5](#_Toc21960)**

[2.1 系统后端处理流程 5](#_Toc15992)

[2.2 系统前端处理流程 5](#_Toc2829)

**[3 系统后端设计 6](#_Toc24594)**

[3.1 图像处理 6](#_Toc1190)

[3.2 文本处理 8](#_Toc28720)

[3.3 文字识别 9](#_Toc5424)

[3.4 结果纠正 13](#_Toc21768)

**[4 系统前端设计 14](#_Toc6121)**

[4.1 设计函数 14](#_Toc23349)

[4.2 设计URL 15](#_Toc29686)

**[5 结果展示 15](#_Toc11897)**

[5.1 首页 15](#_Toc10893)

[5.2 识别结果 15](#_Toc17831)

**[6 方案总结 16](#_Toc20429)**

**[致 谢 16](#_Toc17066)**

**[参考文献 17](#_Toc2836)**

全文共\_\_\_\_\_16\_\_\_\_页\_\_\_\_\_\_5414\_\_\_\_\_\_字

1 绪论

* 1. 研究背景

文字是人们日常交流使用最多的形式之一，随着近些年互联网和智能手机的发展，人们的沟通方式变得多样化，大量的文字变成了语音和图像，面对形式多样的语音数据和图像数据，计算机固有的处理方式逐渐变得低效，如何让计算机看懂图像甚至理解图像是提高信息处理效率的关键点。

Optical Character Recognition(OCR，光学字符识别)是对含有文字内容的图像进行文字提取和数字化的过程。传统OCR识别系统主要应用于从扫描文档中提取文本，其处理流程主要包括:图像预处理、图像二值化、单字符切割、连接识别结果等。

由于中文具有类别多、结构复杂的特点，传统OCR识别系统的单字符切割难以准确识别汉字，例如，＂从＂误分为＂人　人＂；深度学习是让计算机具备和人脑一样的学习能力，得益于深度学习在图像领域的快速发力，Convolutional Neural Networks(CNN，深度卷积神经网络)配合Connectionist Temporal Classification(CTC，连接时序分类)可以实现端到端的不定长文本识别，不仅可以解决传统OCR识别系统的弊端，同时可以获得较高的识别率。

* 1. 研究难点

身份证信息的OCR识别作为成熟的人工智能应用之一，广泛应用于用户注册、银行开户、交通出行等多种场景，大幅提升了信息的处理效率；尽管如此，某些场景下的身份证OCR识别仍然存在着问题，例如，复印件上的身份证信息提取，除光照因素外，还受到复印质量、水印盖章等多种外界因素影响，使得通用身份证识别系统很难达到满意的结果；本赛题提供的数据集相比普通的身份证信息，具有以下特点：（1）图像清晰度参差不齐；（2）图像中的部分要素信息被加盖的水印遮挡；（3）复印件中身份证的位置、方向较为随意；（4）训练集数据未提供各要素的位置信息，只提供所有要素的文本内容。

* 1. 研究内容

针对赛题数据的特点，我提出一套针对身份证复印件信息提取的技术方案，整体思路如下：（1）定位身份证正反面顶点信息，提取身份证区域；（2）根据各要素相对位置提取各要素区域；（3）识别文字并对结果进行矫正，输出最终识别结果。

本方案主要研究内容包括：（1）对复印件图片进行图像切割、图像旋转矫正和去除水印等图像预处理方法；（2）使用模板匹配技术定位身份证各要素信息位置；（3）使用中文合成数据集预训练文字识别模型，在比赛提供的数据集上微调得到最终文字识别模型；（4）使用公开的全国省市区级行政编码纠正签发机关识别结果。

1. 系统流程介绍
   1. 系统后端处理流程

后端系统包括三个模块：（1）图像处理：定位身份证顶点模块；（2）文本处理：定位文本行模块；（3）文字识别：识别文字信息模块；处理流程如图1：

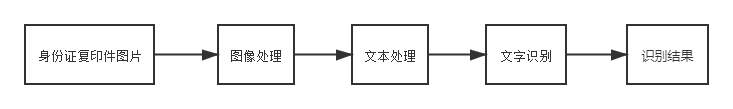


图1 后端处理流程

* 1. 系统前端处理流程

前端系统包括两个界面：（1）WEB端——首页：上传身份证复印件图像；（2）WEB端——结果展示页：显示文字识别结果；处理流程如图2：



图2 前端处理流程

1. 系统后端设计

后端系统采用Python编程语言搭建，Python语言具有简单、速度快和丰富的第三方库等优点，可快速的构建深度学习应用。

* 1. 图像处理

图文识别的成功率一定程度上依赖于图片处理结果的好坏，图片处理的目的在于：剔除图片中的噪声数据，凸显有效识别区域；好的处理结果可以有效提高识别准确率。目前，常用的图像处理技术包括:图像滤波、图像二值化、图像目标检测、图像形态学操作。

* + 1. 图像锐化

图像滤波是图片预处理中常用的方法，它需要在不丢失图像细节特征的条件下对目标图像的噪声进行抑制；关键在于选择一个可以有效消除噪声的滤波器，在选择时需要考虑两点：能有效地去除背景和背景中的噪声，同时，能很好地保护目标图像的形状、大小和独特的几何结构。通过对数据集的分析，图片有不同程度的模糊现象，直接输入模型会难以提取身份证轮廓信息，因此需要采用锐化内核对图片进行滤波操作，消除图片中的模糊现象，使后续身份证的轮廓信息更易于提取。

# 锐化前后对比图

* + 1. 图像二值化

图像二值化是指将灰度图中的像素值设置为0或者255，使整个图片呈现出黑白效果，便于凸显出目标区域的轮廓，同时经过二值化后的图像，其数据量减小，图像变得简单，利于后续的图像处理。

本文二值化方法使用openCV开源库中的adaptive Threshold函数，它是一种局部性的阈值，可以动态自适应地调整属于自己像素点的阈值，相比于全局阈值，会有更好的处理结果；它通过人为设定一个区域，在此区域内比较各像素点与区域里面像素点的平均值的大小关系确定此像素点是属于黑或者白。

* + 1. 图像目标检测和切割

定位身份证顶点的前提是检测身份证的轮廓信息，目标轮廓检测也是图像预处理中常用的技术，本文轮廓检测方法采用openCV开源库中的findContours函数，因为身份证轮廓为矩形且只需要保存矩形的四顶点坐标，在参数上选择cv2.CHAIN\_APPROX\_SIMPLE压缩水平、垂直、对角线方向的元素，只保留该方向的终点坐标。

图像目标检测部分输入二值图，之后进行腐蚀膨胀处理，对检测到的轮廓进行面积筛选，轮廓面积大小占复印件图像的0.05~0.5为身份证轮廓信息，否则为无效信息，得到身份证正反面区域；然后求取该区域的最小外接矩形，计算矩形边水平线夹角，并根据最小外接矩形及其与水平线的夹角使用投影变换。目标检测如图3：

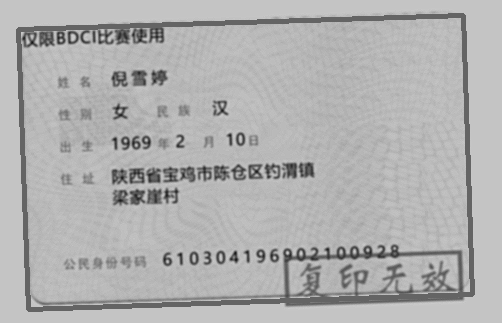


图3目标检测

* + 1. 图像方向矫正

图像匹配技术的目的在于寻找待处理图像中的特定元素，进而确定图像中的水印位置、字符位置．主要技术包括:相似性测度匹配方法、模板匹配方法、尺度不变特征变换，本文采用模板匹配方法。

在身份证的左上角标有＂仅限BDCI比赛使用＂的logo，这可以作为判断身份证方向的标志，首先提取出logo的模板图像，分别将模板图像与身份证的左上角和右下角进行卷积运算，如果右下角的卷积值大于左上角，说明身份证方向倒置，则进行图像翻转操作。

纠正身份证倒置问题后，还需要准确识别出身份证的正面和反面，因为身份证正反面信息内容不同，这里采用模板匹配的方法：模板匹配方法的处理流程是：在待处理图像中进行平滑移动得到匹配目标，此方法要求匹配目标不可旋转，否则无法找到匹配目标；本文数据集中待处理的印章水印和logo位置相对固定，不会影响匹配结果；通过拉普拉斯函数计算得到正反面模板图像，拉普拉斯图像保留了丰富的形状信息，这可以精确地识别身份证正面或背面的要素信息。方向矫正如图4：



图4方向矫正

* + 1. 图像去除水印

图片中的印章水印是此次图文识别任务的一个难点，水印的位置不固定且有一定几率会遮挡到文字部分，影响后续的识别效果；本文采用的方案是先利用模板匹配定位到水印位置，再对水印区域的灰度值进行线性变换以弱化水印。去除水印效果如图5：

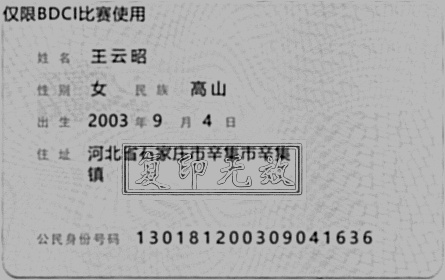


图5去除水印

* + 1. 图像处理流程图

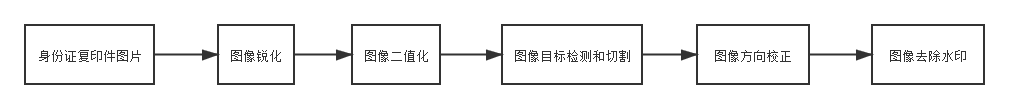


图6图像处理流程

* 1. 文本处理
     1. 切分文本行信息

通过上述处理，得到干净的身份证正反面图像；身份证页面格式固定，通过计算多幅图像得到身份证正反面模板图像，根据模板图像确定各要素区域的相对位置信息和区域长度；这些位置信息可以准确地帮助我锁定图像中各要素的位置范围，提高切分文本行图像的准确性。

将身份证正反面图像与模板图像进行匹配，在匹配过程中，会得到两个值，一个是最高响应值代表两者的匹配程度，另一个是最高响应值的坐标代表目标图像相对于模板的位置偏差，根据得到的位置偏差修正各要素的坐标信息，实现精准切分文本行。

* 1. 文字识别
     1. 深度学习

深度学习是机器学习的分支，由多层感知器构筑起来的深度学习网络可以组合低层特征形成抽象的高层特征，进而识别出数据的分布式特征，实现既定的任务要求。深度学习的原理在于模拟人脑进行分析学习解释数据，它在2012的ImageNet比赛中大放光彩，一举将图像分类的正确率提高了约10%，此后，深度学习被广泛应用在图像和视觉领域，大大超越了传统机器学习所取得成绩。

TensorFlow是由Google主导开发的深度学习框架，是目前主流的深度学习框架之一，它拥有良好的延展性和支持分布式系统，在多GPU和多机上拥有极强的灵活性，随着Google不断更新，现在使用TensorFlow框架的应用可快速部署在各类服务器、PC终端和网页并支持GPU和TPU高性能数据运算。

Torch是由Facebook主导开发的深度学习框架，在Python平台上名为PyTorch，它拥有GPU加速的张量计算和包含自动求导系统的深度神经网络，同时它还有入门简单和快速高效的特性受到广大研究人员的喜爱，本文采用PyTorch深度学习框架。

Caffe是由伯克利大学主导开发的深度学习框架，它具有完全开源的特点，同时提供了用于训练、测试等完整工具包帮助开发者快速入门。

* + 1. 卷积神经网络结构

卷积神经网络(Convolutional Neural Network,简称CNN)，是一种前馈型神经网络结构，主要结构包括卷积层、池化层。

卷积层用于提取输入图像的特征，设置不同的卷积核可以提取不同的特征图，通过多种特征图准确了解输入图像的特征，它的基本计算流程是对输入图像的局部矩阵和卷积核矩阵各个位置上的元素相乘，然后相加所有结果得到卷积层结果。

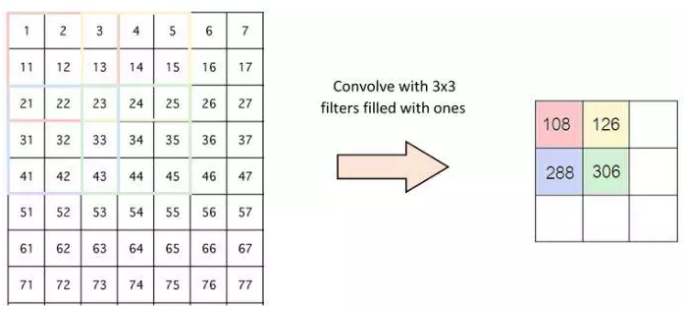


图7卷积层运算过程

池化层又称采样层，用于特征图降维，减少卷积层的参数数量和压缩数据，提高模型的容错性，主要分为最大池化和平均池化。

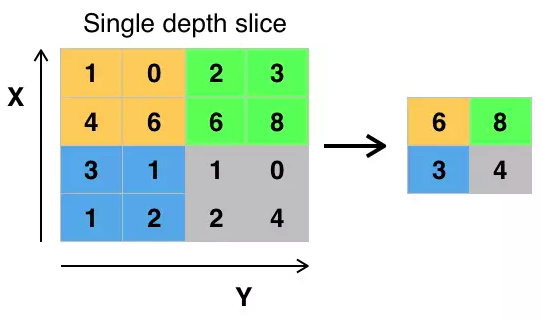


图8池化层运算过程

* + 1. 中文识别模型

输入模型的图像尺寸为1\*22\*220，经过上述处理得到的文本行图像，因为是单行文本且文本长度不固定，所以图像的宽度一定大于高度，经过卷积层计算得到特征图，模型最后使用Bottleneck层对数据进行降维处理，输出类别数\*1\*字符数的特征向量。

Dropout是一类用于神经网络训练或推理的随机化技术，它通过将隐藏的单元激活设置为零，并在训练中设置一定的概率来实现。在第一层使用dropout是为了增强模型的泛化能力；中间各层使用的dropout均为随机屏蔽部分神经单元，解决模型过拟合问题。

Bottleneck层的使用有两个作用：（1）使用1\*1的卷积核对数据进行降维处理，缩短模型训练时间；（2）使用BatchNorm对卷积神经网络进行正则化，同时避免模型训练过程中的梯度消失问题，让训练过程更加稳定，得到更优的模型结果。

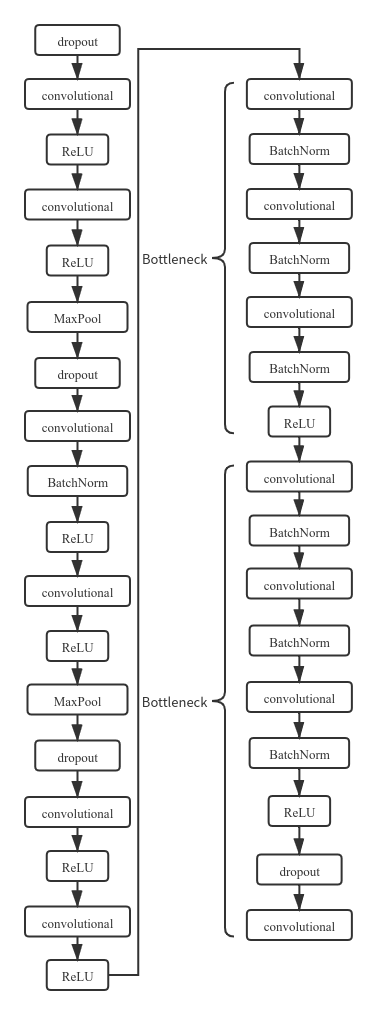


图9中文识别模型

* + 1. 数字识别模型

正文

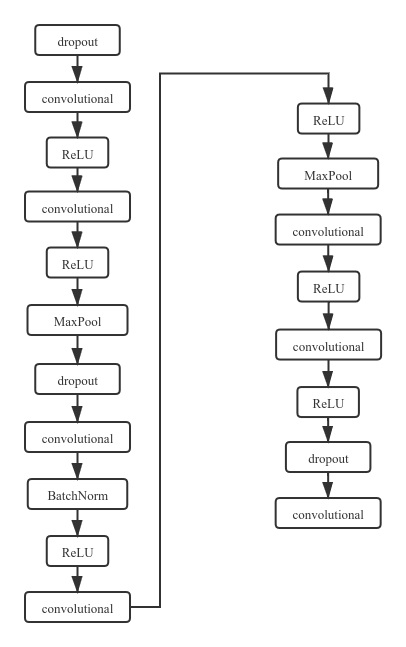


图10数字识别模型

* 1. 结果纠正

针对身份证信息中性别、民族、日期等识别结果都在有限范围的类别内，采用基于词典的Beam Search方法优化识别结果。

生日信息校正：生日本身具备一定的规律，如年份必定是19或20开头，月份和日期有一定的数值范围，身份证号码中的地7-14位与出生日期的年月日是对应一致的，基于这些限制，对生日信息进行校正。

身份证号码的前六位为地址码，因为数字相比中文识别更为准确，将识别出的地址码在region\_code字典中查找得到签证机关的识别结果；相比中文识别模型得到的结果，利用字典确定的签证机关信息更为准确且节省模型运算时间。

有效期校正：有效期和生日日期一样，具备日期约束，我将相似的方法应用于有效期校正；同时，身份证的有效期根据办理时的年龄可以分为四类：五年、十年、二十年、长期。根据这些特点，推算出对应的有效期截至日期，最终实现有效期校正。

1. 系统前端设计

前端系统采用Flask轻量级WEB应用框架，相比同类应用框架，Flask更轻便、更灵活，因此，它又被称为“微框架”：在保持代码简洁的同时快速实现网站的搭建。Flask核心函数包括Werkzeug和Jinja2，Werkzeug支持URL路由请求集成；可以处理HTTP基本事务，快速响应客户端推送过来的访问请求。Jinja2支持自动HTML转移功能；系统运行速度很快，页面加载过程会将源码进行编译形成python字节码，从而实现模板的高效运行。工作流程如图：



图11Flask工作流程

* 1. 设计函数
     1. 预加载模型

识别模型是系统中最重要和最耗时的模块，在WEB服务初始化的过程中进行预加载模型数据，不仅可以在前端网页中提供更快的识别速度，而且可以避免重复加载模型数据节约服务器内存等资源。

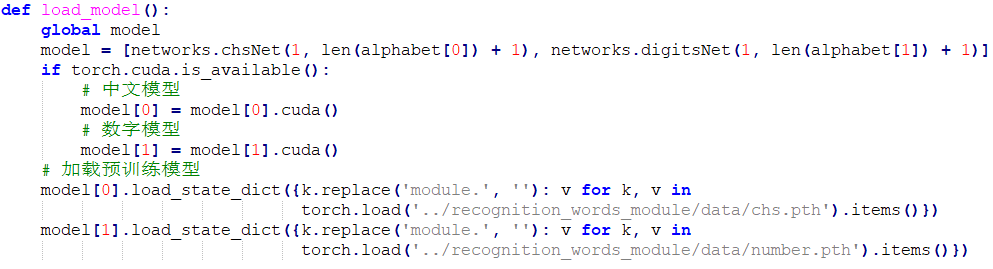


图12预加载模型函数

* + 1. 文件上传

前端系统通过文件上传控件接收图像，当客户端的请求变为POST时，后端系统读取文件对象并判断上传文件是否为图片格式，若是则进行后续图像识别流程。

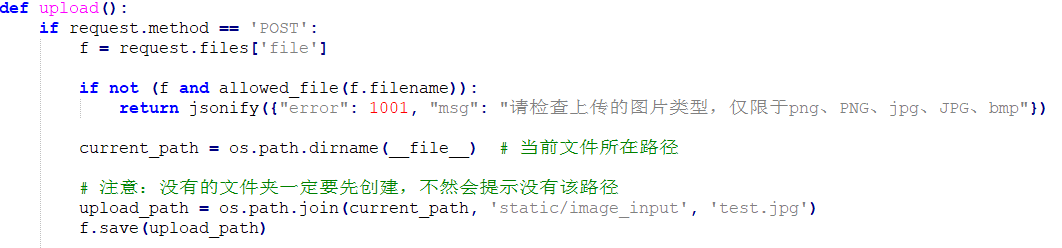


图13文件上传函数

* 1. 设计URL
     1. 首页

首页的背景选自内蒙古师范大学计算机学院官网图片，网页标题采用灰底黑字，文件上传按钮居中显示，使访问者快速获悉网站的主要功能。

* + 1. 结果页

结果页展示身份证复印件识别结果和身份证正反面图片，便于访问者核对结果是否准确。

1. 结果展示
   1. 首页

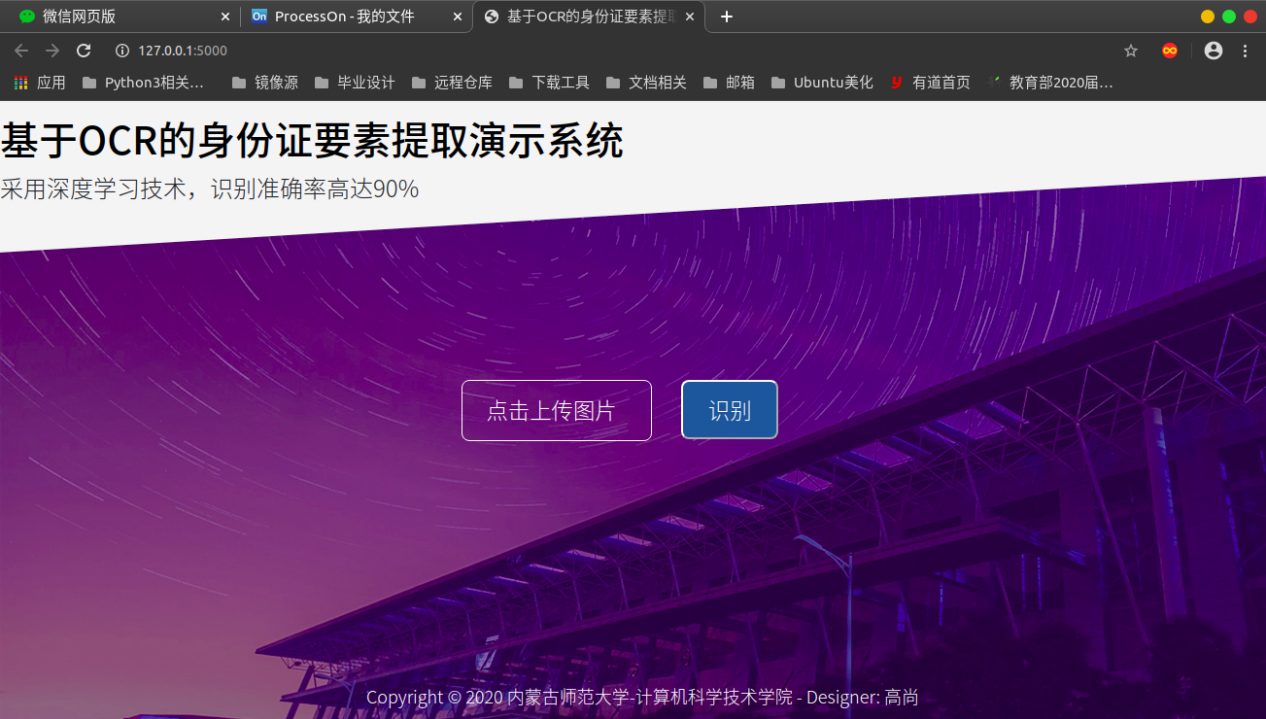


图14首页展示

* 1. 识别结果



图15结果页展示

1. 方案总结

本套技术方案采用CNN+CTC的技术路线，提供了一种端到端的识别方案，旨在解决OCR图像中的身份证信息识别，竞赛方提供的数据集为生成数据，与真实样本存在差异；此方案针对竞赛数据设计实现，故依赖许多强假设，当应用场景发生变化时，需要对方案进行相应的调整。

致 谢

大学四年的学习时光是快乐的，也是短暂的，在这四年中，我不仅收获了丰富的专业知识，也得到了班里同学们和老师们的热心帮助，在此，我向他们表示衷心的感谢。

首先，真诚地感谢我的论文指导老师李艳玲教授，感谢她在我做毕业设计期间给我的耐心指导，指引我在深度学习方向一步步成长。

最后，我要感谢我的家人在我做毕业设计期间对我生活上的帮助和精神上的鼓励，使我以积极的心态面对作品中的种种困难。

参考文献

1. 戈嘉宇,刘为嵩.基于深度学习的身份证识别系统的设计与实现[J].电子世界,2020(02):109.

[2]冯海. 基于深度学习的中文OCR算法与系统实现[D].中国科学院大学(中国科学院深圳先进技术研究院),2019.

[3]尹佳利. 自然场景文字检测与识别研究[D].石家庄铁道大学,2019.