

学校代码: 10246
学 号: 17210240072

復旦大學

硕 士 学 位 论 文
(专业学位)

基于注意力机制的自然场景文字识别
Scene Text Recognition Based on Attention Mechanism

院 系: 计算机科学技术学院
专 业: 计算机技术
姓 名: 褚倩云
指 导 教 师: 薛向阳 教授
完 成 日 期: 2019 年 10 月 8 日

指导小组成员名单

薛向阳 教授

李斌 研究员

金城 教授

张玥杰 教授

目 录

目 录.....	I
摘 要.....	III
Abstract.....	IV
第一章 绪论.....	1
1.1 研究背景及研究意义.....	1
1.2 发展历程及研究现状.....	2
1.3 研究内容与创新点.....	4
1.4 论文结构安排.....	5
第二章 自然场景中的文字识别.....	6
2.1 自然场景文字处理流程.....	6
2.2 自然场景文字图像特点.....	8
2.3 自然场景文字识别技术.....	9
2.3.1 基于字符的识别.....	9
2.3.2 基于单词的识别.....	10
2.4 目前文字识别技术存在的问题.....	11
2.5 常用数据集及评价指标.....	14
2.5.1 常用数据集	14
2.5.2 评价指标	17
2.6 本章小结.....	18
第三章 注意力机制理论.....	19
3.1 注意力机制理论概述.....	19
3.1.1 注意力机制简介.....	19
3.1.2 编码解码注意力.....	20
3.2 注意力机制的模型改进.....	21
3.2.1 在注意力向量的加权求和方式上改进.....	21
3.2.2 在匹配度的计算方式上改进.....	22
3.2.3 其他特殊的注意力.....	22
3.3 注意力机制在场景文字识别领域的应用.....	23
3.4 本章小结.....	25
第四章 基于注意力机制的自然场景文字识别.....	26
4.1 算法框架描述.....	26
4.2 基于 CAN 和 BiLSTM 的图像编码.....	27
4.2.1 方法概述	27
4.2.2 基于 CAN 的空间局部特征提取	28
4.2.3 基于 BiLSTM 的序列上下文特征提取	30
4.3 基于 ASGN 的图像解码.....	32
4.3.1 方法概述	32
4.3.2 ASGN 网络设计.....	32
4.4 系统损失函数设计.....	34
4.5 本章小结.....	34
第五章 实验与分析.....	35
5.1 实验使用数据集介绍.....	35
5.1.1 训练集	35
5.1.2 测试集	35
5.2 图像预处理.....	35

5.3 实验设置.....	36
5.3.1 实验环境	36
5.3.2 训练策略	37
5.4 实验结果.....	38
5.4.1 消融实验	38
5.4.2 模型优化实验	39
5.4.3 标准数据集实验.....	40
5.5 实验分析.....	41
5.5.1 识别过程可视化.....	41
5.5.2 泛化能力分析	43
5.5.3 综合性能分析	44
5.6 中文场景讨论.....	45
5.6.1 实验设置及结果.....	46
5.6.2 实验结果分析	46
5.7 本章小结.....	47
第六章 总结与展望.....	48
6.1 本文总结.....	48
6.2 工作展望.....	48
参考文献.....	49
致谢.....	54

摘要

随着智能终端的发展及互联网技术的普及,在这个人工智能的新时代,越来越多的应用成熟落地,走进人们的生活,这与计算机视觉领域的快速发展不无联系。图像文本识别作为计算机视觉最重要的任务之一,受到学术界和工业界的极大重视。传统的图像文本识别,主要面向质量较高的扫描型文档图像。此种图像排版规整,背景干净,已经能够达到很高的识别水准。但是对于自然环境下拍摄到的文本图片,识别难度却大大增加。图片一旦出现遮挡、模糊、扭曲等情况,识别准确率将迅速下降。因此,针对自然场景下的文字识别技术亟待提高,其在自动驾驶、车牌识别、证件识别、图像理解等领域中有着广阔的前景。

本文针对自然环境这一特殊应用场景,在深入分析的基础上,探究自然场景文字识别的高效解决方案,同时结合深度学习中的注意力机制,研究和设计了一个自然场景文字识别方法,该方法能够更好地捕获图片空间信息以及文本的上下文信息,并在真实复杂场景下取得良好的文本识别结果,主要研究内容包括:

(1) 提出了基于卷积神经网络(CNN)和双向长短期记忆网络(BiLSTM)的二次图像编码方法。在CNN一次编码阶段引入通道注意力机制(Channel Attention Mechanism),通过特征重标定获得更为重要的图像通道特征。利用BiLSTM进行图像二次编码,捕获序列上下文特征。该编码方法强调了图像空间信息,提供了图像级注意力(Image-Level Attention)选择,可以更好地捕捉图像的空间特征及上下文特征,进行了特征增强,使模型具备更强的特征表达能力。

(2) 提出了基于ASGN(Attention-based Sequence Generation Network)的图像解码方法。ASGN通过在门控循环单元(GRU)解码阶段引入注意力机制,对当前时刻神经网络关注点进行建模,同时完成字符定位和识别,提供了文本级的注意力(Context-Level Attention)选择,从而代替主流模型中常见的CTC模型,避免高复杂度的匹配搜索。实验表明该方法具有良好的字符定位和预测能力。

(3) 实现了一套基于层次化注意力机制的自然场景文字识别方法。该方法能够支持多种神经网络结构,并分别在编码阶段和解码阶段引入注意力机制,获得更加有效的特征表达能力和字符预测能力。为了验证该方法的有效性,在IIIT5K、SVT、ICDAR2003和ICDAR2013标准数据集上进行了实验,实验结果表明,本文算法相较于其他算法在识别精度和泛化能力方面有明显提升。

关键词: 注意力机制; 特征编解码; 自然场景; 文字识别

中图分类号: TP391

Abstract

With the rapid development of intelligent terminals and Internet technology, more and more applications have been developed and entered human life in this new era of artificial intelligence, which depends on the rapid growth of computer vision. As one of the most important tasks of computer vision, image text recognition has attracted great attention from both academia and industry. Traditional text recognition technology, mainly serves for high-quality scanned document images. Generally speaking, such images have a neat layout, a clean background, and can be recognized at a high accuracy level. However, for pictures taken in the natural environment, the difficulty of text recognition is greatly increased. For example, when the picture is occluded, blurred, distorted, etc., the recognition accuracy will drop significantly. Thus, the text recognition technology for natural scenes, which has wide prospects in many fields such as automatic driving, license plate recognition, identity recognition, image caption and so on, needs to be improved.

In view of the special application scenario of natural environment, this paper explores the efficient solution of text recognition in natural scenes based on the deep analysis. In this thesis, we study and design a natural scene text recognition method, which utilizes the attention mechanism in deep learning, to better capture the spatial information of the image and the context information of the text meanwhile. In addition, it also achieves good text recognition results in real complex scenes. The main research contents include:

(1) A double image encoding method based on convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) is proposed in this thesis. The channel attention mechanism is introduced in the CNN on the first encoding stage, so that more important image channel features are obtained through feature recalibration. Then, we capture the sequence context features by using BiLSTM for the second encoding stage. The double encoding method emphasizes image spatial information and provides image-level attention. In this way, it better captures the spatial features of the image as well as the context features. This method provides a feature augmentation and makes the model have stronger feature expression capability.

(2) An image decoding method based on ASGN (Attention-based Sequence Generation Network) is proposed in this thesis. It introduces the attention mechanism in the GRU (Gated Recurrent Unit) decoding stage to model the current focus of

neural network and provides a Context-Level Attention. It locates and recognizes characters at the same time and circumvents the use of the CTC model adopted by many state-of-the-art models thus the searching complexity is reduced. Experiments show that the method has good character positioning and prediction ability.

(3) A type of natural scene text recognition method based on the hierarchical attention mechanism is proposed. The method supports a variety of neural network structures, and introduces attention mechanisms in the encoding phase and the decoding phase, respectively, to obtain more effective feature expression ability and character prediction ability. In order to verify the effectiveness of the proposed method, a set of experiments were carried out on the benchmark datasets of IIIT5K, SVT, ICDAR2003 and ICDAR2013. These experimental results illustrate that the proposed method achieves significant improvement in recognition accuracy and generalization ability compared with other methods.

Key Words: Attention Mechanism; Feature Codec; Natural Scene; Text Recognition

CLC Number:TP391

第一章 绪论

1.1 研究背景及研究意义

人类与外界的信息交流，主要有视觉、听觉、嗅觉、触觉等方式。古有“百闻不如一见”的说法，这不仅充分揭示了客观世界的经验规律，也足以证明视觉在人类生活中的重要性。通过视觉，人类可以感知外界物体存在，获得物体的各种必要信息，如大小、形状、色彩等，视觉是人类最复杂、最重要的感官。如何让计算机像人类一样去“看”，能够对外界信息进行获取和感知，这是计算机视觉（Computer Vision）的基本研究目标，其主要任务包括图像分类、图像分割、目标检测、目标识别等，而计算机视觉的终极研究目标则是让计算机像人类一样“思考”，实现对一张图像的理解^[1]。

图像理解是计算机视觉领域中一项极具挑战性的任务。因为普通的视觉元素，如图片中的点、线、面及其相互关系，缺乏足够丰富的上下文约束，从而对一副图像的理解不够全面准确。根据 Marr^[2]在《Vision》中所阐述的视觉计算理论，视觉是一个复杂的信息处理（information processing）任务，是把一些符号表象（representation）变成另一些符号表象的过程。这一过程从外界将光线投射到视网膜上开始，直到大脑皮层形成某种知觉为止，图像可以通过计算从二维平面复原到三维几何结构。这实际上是一种视觉信息的表达和加工。对于图像理解来说，物体的边缘检测是底层的图像理解，物体的语义分割是中层的图像理解。不同于普通的视觉元素，文字元素往往能够表达更为直接和清晰的信息，属于高层视觉元素，因此其所包含的信息也属于高层的图像理解。

另一方面，在信息传播的 4 种方式中，视频和音频在短时间内都无法得到信息的快速分解，只有文字和图片能够将需要表达的信息直观地展示，帮助用户更好地理解图片作品的意图和主旨。在一张有文字的图片上，文字是高层语义信息（high level semantics）的载体，图片是文字内容辅助信息的载体，两者相辅相成，信息互补，因此文字与图片相比单纯的图片浏览更容易让人们记忆与理解。此外，有文字的图片还隐含着丰富的场景信息，需要分析和发掘。例如，针对街道上一个普通的路标指示牌，分析其所描述内容就可以获取很多有用信息。比如可以了解目前个人所在位置，是在哪条街道，哪个路口；指示牌所指示的目的地在什么方向，什么位置；目的地距离目前所在位置有多远等。文字往往蕴含高层语义信息，可以充分挖掘利用其所包含的场景信息。因此，文字识别技术拥有大范围的视觉应用，如基于内容的图像、视频检索，目标定位和跟踪，智能交通导航，车牌证件识别等等。

传统的字符识别，即广义上的光学字符识别，统称为 OCR 技术，它的发展已经趋于成熟，主要针对印刷体字符的扫描型文档，利用摄像头、扫描仪等电子设备，采集纸质文档的图像，再进一步通过识别算法，将图像内容转成计算机文字。经过多年的发展，在图片质量较好的情况下，识别准确率可高达 95% 及以上。但是由于该项技术条件要求较为严苛，需要图片上的文字背景干净，字体单一，排版分布规整，且图片本身分辨率较高，所以一旦涉及到自然场景中的文字识别，效果会大打折扣。这是因为自然场景中的文本图像采集存在很多不确定性因素，不均匀的光照，随机的遮挡和扭曲，一定程度的运动模糊，背景复杂，背景污损等，都会造成非常极端的采集条件，严重影响图片中文本的成像质量。这也对图像文本的识别提出了巨大的挑战。如何在识别过程中解决和应对这些困难，是我们研究的主要目标。

因此，不论是从学术研究上，还是工业需求上，自然场景下的文字识别技术在学术界和工业界已经成为一个热门的研究话题。研究自然场景下的文本识别，突破现有技术的瓶颈，有着显著的理论意义和实际的应用价值。

1.2 发展历程及研究现状

OCR 这个概念最早是由德国科学家 Tausheck 于 1929 年在一项专利中提出，他认为可以利用机器来识别文本字符。随后，美国科学家 Handel 也提出了类似的想法。直到 20 世纪 50 年代，第一个商用的 OCR 产品在美国出现，同时也出现了世界上第一台商用计算机 UNIVAC。到了 60 年代初期，IBM 公司生产出了第一款光学字符阅读器（OCRs）^[3]。随后的 40 多年，西文 OCR 技术得到了突破性进展，广泛应用于商业及工业的各个领域，如识别银行支票、邮政编码、车牌等数字及英文符号，实现了信息的“电子化”处理。商用产品的典型代表之一就是俄罗斯 ABBYY 公司的光学字符识别软件，称为 ABBYY Fine Reader，另一个后起之秀为 Google 公司推出的服务 Google Books，如图 1-1 所示，为 Google Cloud Vision API 的识别结果。



图 1-1 Google Cloud Vision API 的识别结果

字符识别领域的大规模应用得益于 80 年代半导体技术的快速发展，如越来越廉价的兆字节存储器、CCD 图像传感器以及大规模集成电路设计（LSI，Large-scale integration），这使得扫描的图像可以被整页存储起来进一步处理。在 1983 年，第一次出现了手写数字识别系统，随后，日本厂商创造性地将 OCR 引入生产线，可以识别将近 2400 个印刷和手写汉字字符。更多的详细内容，可以查阅文献^[4-5]。这段期间，OCR 技术也迎来了第一次工业浪潮，主要是应用于自动化办公领域。

相比于西文 OCR 的发展，汉字 OCR 的研究是在其基础上发展起来的。在上个世纪 60 年代末期，IBM 公司发表了世界上第一篇关于印刷体汉字识别的论文，虽然只是简单地使用了模版匹配算法，却可识别将近一千个汉字。60 年代末期，OCR 引起了日本学者的极大兴趣，并得到了长足的发展。东芝综合研究所在 1977 年发明了一套可识别两千个单体印刷体汉字字符的识别系统^[6]，日本武藏野电气研究所在 80 年代研制的印刷体汉字识别系统，甚至可识别多达两千三百个多体汉字。

虽然同是拥有方块字的国家，我国直到 20 世纪 70 年代才开始研究 OCR 相关技术的研究。虽然起步晚，但是由于政府的高度重视和大力扶持，我国汉字识别的发展和应用取得了瞩目的成绩。多年来，经过诸多学者的不懈努力和辛勤钻研，我国的汉字识别技术由最初的单体字符识别到多体字符识别，从简单的算法设计、实验验证到成功的商业推广，都印证着我国 OCR 领域的快速发展。我国最早的 OCR 商用产品由南开大学科学家，OCR 核心技术发明人王庆人教授开发并投入使用。我国其他研究机构，高校类如清华大学电子工程系，科研机构如沈阳自动化研究所等多家单位都对 OCR 领域进行了相关产品研发。其中最具代表性的就是清华大学和北京文通公司联合研制的文通 TH-OCR 以及汉王集团开发

的尚书 OCR 系统，不仅占据着市场大部分的份额，也是印刷体汉字识别技术的一座里程碑。

近些年的 OCR 研究热点主要集中在自然场景图像识别以及视频图像方面。最初的图像文本提取与识别研究源于 Ohya 和 Shio^[7]等人对一些场景图片，如路牌、车牌、商店标牌的字符识别，随后 Lee 和 Kankanhalli^[8]进行了对货运集装箱的文本定位与自动提取，Zhong^[9]等针对复杂的彩色图片场景，对 CD 和书籍封面的文本进行了定位和识别，Zhou^[10]等研究了从互联网上采集图像的文本提取，首次利用色彩空间信息进行分类，实现了从传统的灰度图像处理到彩色图像的过渡。为了促进自然场景文本图片的技术研究，素有文档图像识别领域“奥斯卡”之称的国际文档分析识别大会（ICDAR），同时也是文档分析与识别领域最权威的国际学术组织，从 1991 年起，每两年举办一次鲁棒文本阅读竞赛，该竞赛成为检测和评估自然场景、网络图片等文本检测、定位与识别最盛大的国际赛事，也是业内公认的最权威标准，对自然场景文本识别技术的革新与发展起到了极大的推动作用。我们现在正面临着 OCR 领域的第二波浪潮，互联网技术的发展在一定程度上对 OCR 产生了冲击，OCR 技术是否适应新的环境，是否有新的需求，是否需要进一步革新，是我们需要深入探索和亟待解决的问题。

1.3 研究内容与创新点

本文主要研究内容集中在自然场景这一复杂环境中的文字识别，在现有主流文字识别算法的基础上，结合近年来的深度学习前沿技术，提出了一套新型的基于注意力机制的自然场景文字识别方法。本文提出的创新点可概述如下：

（1）提出一种基于注意力机制的图像二次编码方法，网络主要由 Attention-based CNN 和 BiLSTM 构成，在一次编码阶段，CNN 引入注意力机制，构成 Image-Level Attention，可以捕捉到更有利于文本预测识别的图像局部空间特征，在二次编码阶段 BiLSTM 能够编码完成序列上下文特征。采用二次编码的方式进行了特征增强，因此模型具备更强的特征表达能力。

（2）提出一种基于注意力机制的图像解码方法 ASGN（Attention-based Sequence Generation Network）。网络主要由 Attention-Based 的 RNN 构成，提供了 Context-Level Attention，能够对当前时刻神经网络的关注点进行建模，代替了主流方法中的 CTC 模型，在减少预测复杂度的同时，能够更为快速、准确地解码输出。

（3）针对一般情况下的自然场景文本图片，提出了一种通用的层次化注意力识别网络 HARN（Hierarchy Attention Recognition Network），并给出了识别方

法。该方法在多个实验及公开数据集上的结果表明，相较于其他方法，本文方法在识别精度和泛化能力方面有明显提高。

1.4 论文结构安排

本文围绕自然场景下的文字识别技术展开论述，结合深度学习中的注意力机制，介绍了相关技术，详细阐明了提出的创新算法，展示了实验结果，并对结果做了详尽的分析说明。总体来说，本文共分为 6 个章节，安排如下：

第 1 章：绪论。介绍了本文的选题背景，阐述本课题研究意义，然后对 OCR 整个发展历程做出概述，并介绍了本文的研究内容、创新之处和全文的章节安排。

第 2 章：自然场景中的文字识别。首先概述了自然场景文字处理流程，然后分析了自然场景中的文字图像特点，对现今主流的多种文字识别技术进行了分类梳理，指出存在的一些技术难点，并总结了自然场景文字识别的困难和挑战。

第 3 章：注意力机制理论。首先介绍了注意力机制理论，然后说明了注意力机制模型的改进，最后介绍了注意力机制在场景文字识别领域的应用。

第 4 章：基于注意力机制的自然场景文字识别。本章详细阐述了本文提出的基于注意力机制的创新性算法，首先做出了算法框架的整体性描述，然后详细介绍了各个模块和网络结构的设计细节。

第 5 章：实验过程与分析。本章首先简要说明了实验过程中数据集的设置与选取，并介绍了数据集图片预处理的过程。然后分别阐述了实验所需环境以及实验中使用的训练策略，最后展示了实验效果并给出了详细解析。

第 6 章：总结与展望。对本文在自然场景文字识别技术中提出的创新点和所做贡献进行了归纳总结，并对下一步的研究方向做了展望。

第二章 自然场景中的文字识别

文字识别理论自提出以来，经过多年发展，已经取得了瞩目的成绩，并成功应用于各大商业、工业领域。但是随着现实场景下的文字识别需求越来越迫切，原来识别简单背景、单一字体的传统印刷体识别技术已经不再适用，因此研究学者们开始转战自然场景下的文字识别领域，力求突破。

2.1 自然场景文字处理流程

自然场景文字处理流程由两个部分组成：一是文字检测与分割，二是文字识别。文字检测即在一幅图像中将含有文字的区域检测到并正确分割出来，识别过程即对分割出来的文字图像进行一定的处理，如图像去噪、图像增强等，然后正确分类识别。本文主要研究内容为后者，专注于文字识别部分。当然，现在也有研究学者将两者结合起来，形成一个端到端（End-to-End）的统一框架。典型的文字识别流程图如图 2-1 所示。

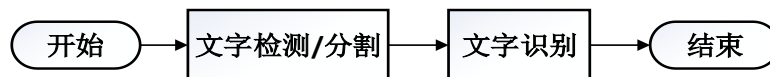


图 2-1 典型的文字识别流程图

传统的文字检测的方法主要沿用两大技术路线。一种是基于滑动窗口方法^[11-12]，设置不同尺度的窗口大小在原图上滑动，将每个窗口所覆盖的区域视为文本候选区，然后提取手工设计的特征（Handcraft Features），利用训练好的分类模型获取该区域的置信度，与设定的阈值比较后判断每个窗口中是否含有目标文字，检测结果再使用非极大值抑制（NMS）得到最终结果。该类方法在检测小尺度文本或对比度欠佳的样本比较有优势，能有效避免相邻文本之间的重叠（overlap）现象。第二种是基于连通分量方法，其又可分为边缘检测方法和文本级检测方法。基于边缘检测^[14]的方法首先利用边缘信息，根据框中的轮廓个数与框边缘重叠的轮廓个数，给每个框打分排序，最后进行检测。文献^[15]提取了水平、垂直、左上、右下四个方向的边缘特征，然后采用 K-means 聚类方法得到文本检测结果。文本级检测方法则利用自然场景文本图像颜色相近、灰度值相近等特点，获得待检测图像中的连通区域，进而获得文本候选区域。

文字识别十分依赖分类器，整个识别过程就是对图像文本进行分类的过程。传统的文本分类主要是基于模式识别理论，大体上可以总结为两种。一种是基于结构特征的分类器，通过分析不同粒度的特征信息形成的图像特征信息结构，形成基于结构特征的分类器，但是这种分类器忽略了图像隐含的信息，结构相似的图像并不一定属于同一类别，所以有其局限性。二是支持向量机分类器，将文本

分类问题转化为多分类问题，构建多分类向量机（MSVM, Multi-Class SVM）。但是随着深度学习领域的发展和神经网络的兴起，打破了传统的模式识别方法。神经网络拥有高效并行处理数据的能力，并且具备很强的自学习性，很大程度上避免了繁琐的特征工程。同时神经网络以数据为驱动，只需要提供大量数据即可。因此，使用基于神经网络的文字识别方法可以一定程度上提高识别准确率，而且训练过程较为简单，可以识别比传统的文本分类器更多的字符。

文字识别的处理过程又可以细分为 2-2 所示的流程。其主要功能是能够在分割出来的文本图像上正确识别出文本内容，接下来将对每个步骤进行详细介绍。

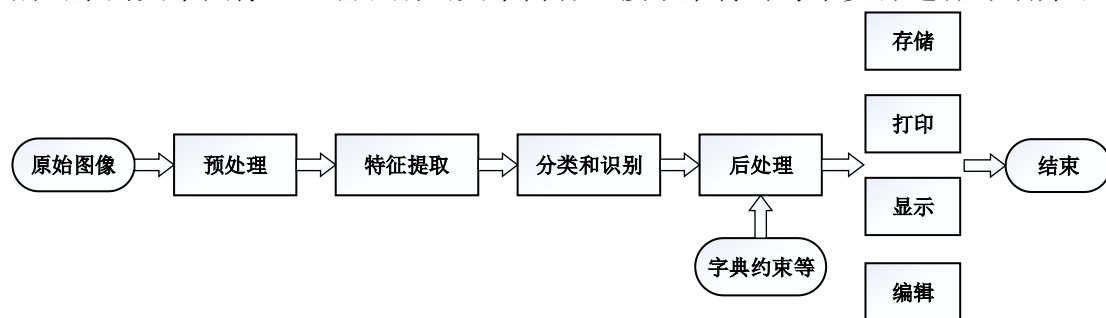


图 2-2 通用文字处理流程

（1）预处理：经检测得到的结果往往包含噪声等影响识别的因素，预处理的目的是减少图像背景的干扰噪声等无用信息，增强有用信息即突出文字部分。通常包括图像灰度化、降噪、二值化。如果文字行有倾斜，还要倾斜检测与校正。最后进行平滑和归一化，将单个文字统一到同样的尺寸，便于后续特征提取和识别。

（2）特征提取：特征提取是整个识别过程最关键的一环，特征信息的优劣直接影响着识别性能。英文和阿拉伯数字的提取比较容易，但是由于汉字构造复杂，字符集数量大，一级常用字符就有 3755 个，因此特征提取比较困难。常见的提取特征包括边缘特征、结构特征、方向特征、笔画特征等。提取了合适的特征之后，有时候还需要特征降维，这是因为特征维数太高，分类器的识别速率会受影响。因此如何更好的降维也是一个研究方向，既能减少维度，又要能保留字符的特定信息。

（3）分类和识别：文本分类需要设计分类器来识别。图像特征被提取之后，再送给分类器进行分类。每个字符都表示一个类，分类器输入提取的特征，输出对应的字符或者单词。

（4）后处理：后处理主要有两个方面，一是版面恢复，因为图像中的文本存在排版、字体大小不一致等情况，最终识别结果则可以通过后处理这一步骤进行规范化。另一个方面是识别校正，对于一些分类器分类错误的字符，可以通过语义模型进行纠正。字典约束通常是一个可靠的方式，例如分类器识别结果为“太

数据”，语义纠正后为“大数据”。当然，对于一些无约束的文字识别，就不需要设计字典约束来对识别结果进行后处理了。

2.2 自然场景文字图像特点

针对自然环境这一复杂场景，文字的识别更为困难重重。这是由自然场景下文字图像的特点决定的。图 2-3 所示为扫描型文本图片，与自然场景文本图片相比，背景干净，字体简单，排布规整，因此在一定条件下，识别效率会很高。而自然场景下的文字图像，具备更多的不确定性和多样性，由于尺寸大小、拍摄风格、文字朝向的差异，以及低对比度和复杂背景而导致的文本变化使得自然场景下的文字识别极具挑战性。

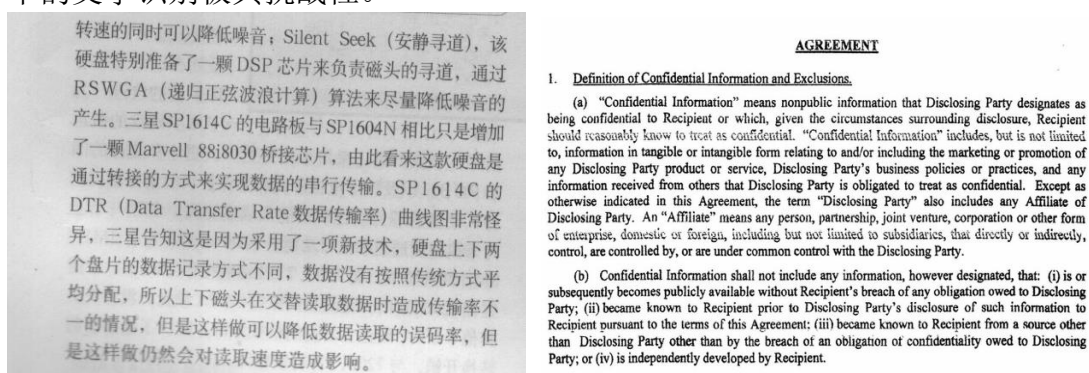


图 2-3 扫描型文本图片

具体来说，自然场景下文字图像的视觉特性有以下几个属性：

(1) 几何属性

自然场景下的文本图像除了自身的尺寸大小等属性，还会由于获取的方式、视角的变化、拍摄的光照、镜头本身缺陷等因素，造成图像中的文本发生几何畸变，为字符识别增加了难度。

(2) 对比度属性

文档图像的背景相对来说比较简单，而自然场景文字图像相对其背景来说，其背景比较复杂，字符区域和背景像素往往会有重叠，灰度信息和颜色信息有明显的对比度，背景颜色和亮度呈现不规则变化。如何将文字从复杂多变的背景中分割出来也成为文字识别领域一个不小的难题。

(3) 梯度边缘属性

自然场景文本图片常常有复杂的空间结构，一方面文本区域的边缘往往较为密集，另一方面图片的梯度信息明显。充分利用这两种属性，可以更为方便地设计和提取特征。

（4）像素属性

传统的扫描型文档图像通常为二值化图像，易于分割。自然场景下的图片通常为三通道的 RGB 图像，色彩千变万化，文字和背景的像素都是多值的，其颜色和灰度分布不均匀，所以仅仅从像素上分离文本和背景十分困难。

（5）分辨率属性

传统文档图像通常拥有大于 300 dpi (Dots Per Inch) 的分辨率，分辨率越高，包含的细节就越丰富。自然场景下的图片由于拍摄手法，地点，光照等不可避免的因素影响，获取到的图片像素往往不高，极端情况下甚至只有 10×10 的分辨率大小，这对字符区域的提取和文本的识别带来极大的困难，所以近些年来研究超分辨率图像也是一个热点话题。

（6）上下文结构属性

文档图像的同一行文本区域，相邻文本的字体大小、笔画宽度、像素值都相近，其版面都是人为设计，布局统一，格式固定，识别起来较为容易。自然场景的文本是对拍摄图片的一种信息补充，如街道的路牌、横幅的标语、商店广告语等，因此字符排列形式多变，彼此之间没有格式对齐的要求，缺乏结构信息，上下文结构属性相对较弱。

基于上述特点，传统的 OCR 技术无法很好地适用于自然场景下的文字图像识别。

2.3 自然场景文字识别技术

2.3.1 基于字符的识别

顾名思义，基于字符的识别即逐个地识别单个字符，再将字符组合起来输出最后的结果，一般经历字符检测或分割、字符识别、字符重组 3 个步骤。

在一些非神经网络的传统场景文本识别算法中，Yao 等人^[16]使用中层特征进行场景文本识别，利用中层特征“Srokelets”来定位字符，同时提取霍夫（HOG）特征，在两种特征的基础上再使用随机森林进行字符的识别。Wang^[12]利用条件随机场（CRF）来模拟字符在空间和语义上的关系以及识别置信度，并借助图结构来完成单词的检测和识别。

神经网络流行至今，也衍生了很多经典算法，基于神经网络的识别算法大多数都使用卷积神经网络作为字符分类器。Alsharif 等^[17]综合隐马尔科夫模型（HMM）和 Maxout 网络^[18]，切割后的字符作为 Maxout 网络的输入，构建了一个包含分割、矫正、识别的端到端系统。Bissacco^[19]等人提出了至今影响力都很大的 PhotoOCR 系统。该系统提取文本图片的霍夫特征，然后使用神经网络作

为分类器，结合 N 元语言模型（ N -gram）的 Beam 搜索获得得分最高的路径，也就是候选字符组合。Jaderberg 等人^[20]通过同样采用 CNN 作为字符分类器，对整张图使用像素级的滑动窗口扫描，结合 Maxout 模型最后分析得出识别结果。

基于字符的识别算法灵活性强，不受字符的排列顺序和方式影响，可以识别任意长度和数量的字符。但其缺点也十分明显，该类方法依赖字符分类器，需要分别设计检测器和识别器，英文等拉丁字母以及阿拉伯数字等类别较少的字符识别较为适合，对于汉字这样复杂庞大的系统就显得力不从心了。此外，由于该类方法的模块较多，需要各模块联合优化，同时需要大量的字符级数据标注，所以整个系统的训练耗时耗力。

2.3.2 基于单词的识别

基于整个单词的识别跳过了字符级检测和分割模块，直接将单词作为一个整体去识别。首先从图片全局提取特征，并表示成向量形式，最后通过回归过程使全局特征回归到目标单词向量。

同是与 Yao 等人^[16]一样通过提取中层特征来识别文字，Gordo 等人^[21]采用的是整个单词的识别，文中的训练数据集进行了字符级的人工标注，提取的中间层特征可以建立整个单词图像的表达。Mishra^[22]等人提出了一个包含了高阶统计语言模型的框架，引入了一个带有字符级标注的大型单词识别数据集，通过构建图结构来推导整个单词。Novikova^[23]等人在文中提出了一种新的自然场景下文本识别任务模型，该模型同时模拟概率模型中每个单词语言的一致性及其属性（如字体和颜色）的一致性，结合了局部似然和位置一致性先验。Goel^[24]等人首先使用基于梯度的特征来表征场景文本图像，并预先制作好其对应的单词合成图像。然后通过将真实图片和合成图像特征比对，利用动态 k 近邻方式来匹配识别文本内容。Rodriguez^[25]等人将单词标签和单词图像嵌入到共同的欧几里德空间中，将识别问题转化为图像检索问题，使用集成的 Fisher 向量^[26]以及结构化 SVM 对图片和单词之间的关系进行编码。Almazan^[27]等人则提出了 PHOC 描述子（称为 label embedding）来表征一个单词，并对整张图像编码得到类似 PHOC 的属性（称为 attribute embedding）。通过这种方式把图片和字符串映射到同一空间进行距离运算。

基于深度学习的方法中，代表性工作有 Goodfellow^[28]等人提出的直接使用深度卷积神经网络在整张图像像素上进行编码，并创建了多个对位置信息敏感的字符分类器识别字符。该方法在 SVHN（Street View House Number）数据集以及验证码识别任务（reCAPTCHA）上取得了巨大的成功。Jaderberg 等人^[29-30]对以上工作进行了细微的改进，用深度卷积网络直接对单词进行分类，输入是整张图

片,输出是类别概率,取最高概率作为对应的字符类别。而且不再预测字符长度,改为在文字结尾处引入了结束位标志符。同时他们也开源了一个大约 9 万常用英文单词的合成数据集,证明了由这些训练数据训练出的模型也能很好的应用到真实世界的文字识别任务。

2.4 目前文字识别技术存在的问题

研究学者在自然场景下的文字识别领域已经积累了丰硕的研究成果,也提出了一系列识别算法,着实解决了部分问题。但是从实际实验结果和工业界需求来看,文字识别技术仍然存在一些问题亟待解决。

(1) 对字符形变十分敏感

自然场景下因拍摄手法、角度等因素导致拍摄出的图片存在透视、仿射等角度变化,另一方面例如商店横幅、标语等由于风力、人为等原因导致文本存在一定的扭曲,这些现象都对文字识别提出了巨大的挑战。而目前提出的主流的文字识别算法均是针对水平文本图像的识别,虽然可以容忍小范围的字符笔画变化,但是不能够处理严重的字符形变问题,性能会显著下降。对此,文献^[31-35]提出了专门针对形变字符文本的识别方法,虽然在一定程度上提高了识别精度,但是模型结构往往比较复杂,而且需要巨大的数据量驱动。图 2-4 展示了一些形变文本图片。



图 2-4 形变文本图片

(2) 难以处理复杂背景

由于自然场景这一特殊应用条件,必然会存在复杂的图像背景,例如遮挡、阴影、污损、文本重叠等干扰比较严重的情况。传统 OCR 对扫描型文档这种背景干净简单的图像识别率已经达到较高水平了,但是应对自然场景下的复杂背景图像还远远不够。一般情况下,复杂背景比较直观的处理方式就是将字符与背景分离,但是这并不容易。2009 年的 ICDAR 鲁棒文本阅读竞赛主要针对这一现象组织了文本图像二值化分割比赛,希望能够促进相关算法的研究。图 2-5 展示了一些复杂背景文本图片,图中的样本存在文本重叠、阴影以及背景污损的情况。



图 2-5 复杂背景文本图片

(3) 对低质量图片识别性能不佳

低质量图片包括图片分辨率较低、字符和背景对比度较低、因拍摄过程中未对焦形成的对焦模糊或拍摄过程中移动拍摄端而造成的运动模糊等情况。由于自然场景文字图像的特殊性，对低质量字符识别的研究工作不够完善。Cheng 等人^[36]针对自然场景下低质量字符识别存在的注意力漂移问题展开了研究，提出了 FAN 方法，该方法用聚焦注意力机制来自动拉回漂移的注意力。图 2-6 展示了一些低质量文本图片。

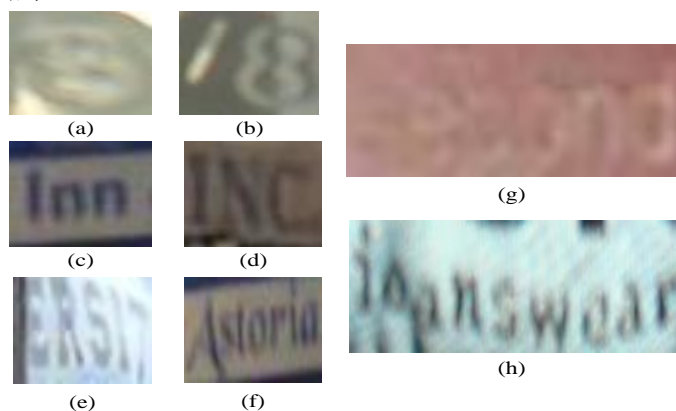


图 2-6 低质量文本图片

(4) 难以捕捉长文本的上下文信息

早期的文字识别方法通常使用分割的手段将字符单个分割进而识别。然而这种分割的方法在一些复杂的情况下很不适用，另外，字符和字符之间存在着一定的依赖关系，如果强行分割，会丢失上下文信息。后期出现的整个单词识别方法能够有效利用字符间的上下文依赖性，所以一定程度上提高了识别准确率，但是对于长文本的识别，还有待提高。如图 2-7 所示，当字符长度超过一定限度，如 20 个之后，就很难捕捉到上下文信息，识别性能会显著下降。



图 2-7 长文本图片

(5) 文本种类繁多

自然场景下的文本多种多样,除了常见的 26 个英文字母和 10 个阿拉伯数字,还有汉语、韩语、日语、法语、德语等一些其他的语种。不同的语种,字符呈现的视觉特征差异性很大,这就面临着文字识别中的多语种问题。而且种类繁多的语种自身也存在着不同的文本类别。

另一方面,文本的呈现方式也各不相同,一些商店的广告、标牌往往是手写字体,一些艺术字体的引入也为文字识别增加了不小难度。艺术字体和手写字体的识别的视觉特征与普通文本有很大差异,是自然场景下文字识别的又一研究热点。图 2-8 展示了多种多样的文本图片。



(a) 多语种

(b) 手写体

(c) 艺术字

图 2-8 多种多样的文本图片

(6) 需要样本增广

现阶段比较主流的识别算法均靠数据驱动,它们普遍依赖巨大的数据量以及详细准确的人工标注。对于一些字符分割识别算法,甚至要精确到每个字符的位置,从而能够更好的切割出单字符,并结合单字符样本和多字符样本联合训练。在一些算法的训练中还需要制作包含文字的正负样本来进行文字区域非文字区域的检测识别。为了提高算法识别性能,样本的制作都要尽量贴近真实场景,并尽可能增加其多样性,这对样本的制作和获取提出了不小的挑战。

综上所述,自然场景下的文字识别技术仍然存在一系列不容忽视的问题,这些问题是影响识别性能好坏的直接因素,需要引起足够的重视,在深入分析以上问题后寻找合适的解决途径。

2.5 常用数据集及评价指标

2.5.1 常用数据集

文字识别数据集根据图像采集方式可划分为三类：自然环境下采集的文本图像数据集；手写文本图像数据集；计算机合成的文本图像数据集。每种数据集都对应不同的OCR识别处理方法。一方面，找到研究领域合适的数据集，并将自己的方法应用其中，是至关重要的。另一方面，数据集往往充当基准（benchmark）的角色，算法在公开数据集上达到一定的性能效果将更具说服力和权威性。接下来简单介绍场景文字识别常用数据集。

（1）ICDAR2003^[37]

ICDAR2003 数据集于 2003 年发布，作为当年 Robust Reading Competition 的标准数据集，包括自然场景文本图片和人工合成图片。数据集文本朝向主要以水平为主，包括训练集图片 1157 张，测试集图片 1111 张以及用作样例的图片 171 张。该数据集同时包括单词级别的图片以及字符级别的图片，如图 2-9 所示，分别是单词图片和裁剪出来的单个字符图片。



图 2-9 ICDAR2003 数据集

（2）ICDAR2011^[38]

ICDAR2011 数据集于 2011 年发布，并作为当年 Robust Reading Competition 的标准数据集，包括自然场景文本图片和人工合成图片。它在原来 ICDAR2003 数据集的基础上扩大，并修改了 ICDAR2003 的一些不足之处，例如补全了丢失的 ground truth 信息，bounding boxes 也与文本之间更加紧密等。数据集共包含 1564 张裁剪后的单词图片。如图 2-10 所示，为 ICDAR2011 图片示例。



图 2-10 ICDAR2011 数据集

(3) ICDAR2013^[39]

ICDAR2013 又称 Focused Scene Text Challenge，数据集在 2013 年发布，作为 2013 年的 Robust Reading Competition 的标准数据集。图片样本多为水平的场景文本字符，存在一定程度的小角度倾斜，同时包含已经裁剪过的字符级图片和单词级图片，包括训练集 3567 张，测试集 1439 张。文字总体来说较为清晰，只有英文文本。如图 2-11 所示，为 ICDAR2013 图片示例。



图 2-11 ICDAR2013 数据集

(4) IIIT-5K^[22]

IIIT 5K 单词数据集是从 Google 图像搜索中获取的，包括广告牌，商店招牌，门牌号，房屋铭牌，电影海报等。该数据集包含来自场景文本和人工合成的共 5000 个裁剪之后的单词图像，每张图像分别对应一个 50 词和一个 1000 词的词典。数据集分为训练和测试部分，可用于 large-lexicon 裁剪后的单词识别。此外，该数据集还提供了一个包含超过 50 万个单词的词典。如图 2-12 所示，为数据集的部分样例。



图 2-12 IIIT-5K 数据集

(5) MSRA TD-500^[40]

MSRA TD-500 是一个多方向文本的数据集，大部分文本都选自导向牌之类的标志。图片分辨率在 1296x864 到 1920x1280 之间，包含中英文，总共 500 张自然场景图片。其中，训练集为 300 张，测试集为 200 张，标注以行为单位，而不是单词，每张图片都完全标注，对于部分难以识别的图像有 *difficult* 的标注。如图 2-13 为数据集的部分样例。



图 2-13 MSRA TD-500 数据集

(6) COCO-Text^[41]

COCO-Text 是一个用于自然图像中文本检测和识别的大规模数据集，每张图片包含多个文本实例和边界框，共有 63686 张图片，分为 3 个细粒度文本属性，其中 43686 张作为训练数据，其余的 2 万作为测试数据，共包含 145859 个文本实例。文本实例有多个种类，可分为机器打印文本与手写文本、易读与非易读文本，以及英文文本与非英文文本。如图 2-14 为数据集的部分样例。



图 2-14 COCO-Text 数据集

2.5.2 评价指标

文本识别领域的评价指标主要有两种，分别是全匹配以及编辑距离，接下做简单说明。

（1）全匹配

全匹配即一个字符串中的每一个字符都识别正确，只要有一个字符识别错误，则认为匹配失败，识别错误样本加一。因此，全匹配又等价于准确率（accuracy），准确率越高，分类器效果越好。假定分类目标只有两类，分别定义成正例（positive）和负例（negative），那么：

True Positive（真正，TP）：将正类正确地判定为正类。

True Negative（真负，TN）：将负类正确地判定为负类。

False Positive（假正，FP）：将负类错误地判定为正类。

False Negative（假负，FN）：将正类错误地判定为负类。

则准确率计算如公式 2.1 所示：

$$accuracy = \frac{TP + TN}{P + N} \quad (2.1)$$

（2）编辑距离

编辑距离（Edit Distance）的概念由俄罗斯科学家 Vladimir Levenshtein 提出，所以又称“莱文斯坦距离”，是用来衡量两个序列之间的相似度。通俗的来说，一个字符串中的部分字符识别正确，可以通过计算编辑距离来衡量算法优劣。编辑距离越小，分类器效果越好。

例如，对于字符串（ A, B ）， k 为 A 通过编辑操作转换成 B 的次数， k 取最小值，表示该字符串之间的编辑距离。编辑操作共有插入，删除和替换三种。假设 A_i 和 B_j 分别为字符串 A 、 B 的前 i 、 j 个字符组成的子串，则计算

$A_i:A[1], A[2], \dots, A[i-1], A[i]$ 转换成 $B_j:B[1], B[2], \dots, B[j-1], B[j]$ 转的最少编辑次数。状态转移方程如公式2.2所示：

$$lev_{a,b}(i, j) = \begin{cases} \max(i, j), & \text{if } \min(i, j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1, j) + 1 \\ lev_{a,b}(i, j-1) + 1 \\ lev_{a,b}(i-1, j-1) + 1_{a_i \neq b_j} \end{cases} & , \text{otherwise} \end{cases} \quad (2.2)$$

其中， a 和 b 分别为字符串 A 和 B 的长度。

2.6 本章小结

本章主要介绍了自然场景中的文字识别相关知识，首先阐述了自然场景文字识别处理流程，文字识别过程包括检测和识别两个步骤，本文主要研究作为识别部分。然后介绍了自然场景中的文字图像特点，简述了一些自然场景中的文字识别技术，常见的文字识别技术可大致划分为单个字符识别和整个单词识别，两种方式各有优劣。接着又深入分析了目前文字识别技术普遍存在的问题，对图像字符的形变十分敏感，对复杂背景和低质量图片文字的识别性能显著下降，而且难以捕捉长文本的上下文依赖性，另外还存在多语种、多文本形式识别性能不佳的问题。这些问题应该引起研究人员的足够关注。最后介绍了文字识别领域几种常用的数据集和评价指标。检测与识别的评价指标差异性较大，识别领域的评价指标主要是全匹配和编辑距离。

第三章 注意力机制理论

3.1 注意力机制理论概述

3.1.1 注意力机制简介

注意力机制（attention）借鉴了人类在观看图片、文本等事物时使用的注意力机制。这种机制是人类视觉的一种选择性处理机制，高效快速地提高了人脑对复杂繁多的视觉信息处理能力。图 3-1 形象地展示了人类的视觉注意力机制，在看到一张婴儿图片时，会将注意力更多聚焦在脸部，在阅读一篇文章时会将注意力更多聚焦在标题和首句。本质上来说，深度学习中的注意力机制类似于人类视觉注意力机制，其目的都是从大量信息中找到最关键的目标信息。



图 3-1 人类的视觉注意力机制^[42]

注意力机制起源于视觉图像，Google Mind 在 2014 年提出的图像分类任务^[43]中首先使用 attention，取得了很好的效果。他们的研究动机也是受到人类注意力机制的启发，在传统 RNN 中引入 attention 机制，通过 attention 去学习一幅图像需要关注的部分，从而减少需要处理的像素和任务复杂度。随后自然语言处理（NLP）领域开始引入 attention 机制，NLP 领域的第一个 attention 任务是 Bahdanau^[44]等人进行的机器翻译（NMT, Neural Machine translation）任务。他们发现神经网络对于短的语句能够较好的翻译，一旦句子增长到某种程度，性能会迅速下降，因此在解码器中使用 attention 机制解决此问题。

3.1.2 编码解码注意力

如图 3-2 所示，注意力机制通常情况下会设计成编码器-解码器结构，但不局限于此，图中 (X_1, X_2, X_3, X_4) 为输入， (Y_1, Y_2, Y_3) 为预测输出， (C_1, C_2, C_3) 为中间语义编码。编码器一般由卷积神经网络或者循环神经网络构成，图像首先通过编码器提取特征，每个特征向量对应原图某块区域。解码时，attention 首先会计算对齐因子，对其因子即为注意力权重，通过对齐因子计算中间语义向量，使注意力区域能够对齐到对应位置，最后使用 RNN 输出预测。这里的对齐可以理解成模型引入注意力机制之后，更为关注当前的输入部分，没有使用注意力的模型又可称为“分心模型”，关注更为平均。

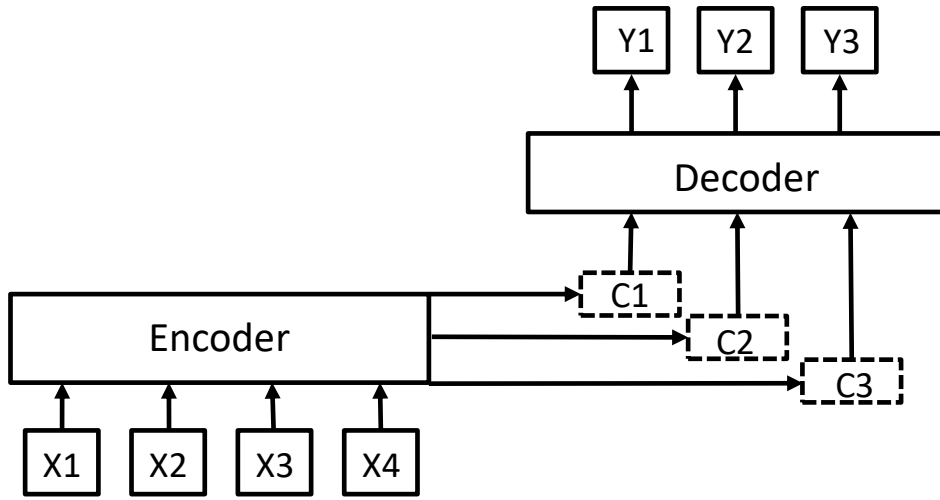


图 3-2 引入注意力机制的编码器-解码器框架

在编码阶段，输入对象通过一个网络模型，如 CNN、RNN、DNN 等，编码为一个向量，这里以 RNN 为例，保留每个 RNN 单元的隐藏层状态，得到 $(h_1, h_2, \dots, h_N) \in R^d$ ， N 为输入句子的长度， d 为编码端隐藏层单元数。

在解码阶段，也以 RNN 为例，对于每一个时间步（timestep），根据解码器的输入和上一步隐藏状态的输出，得到当前步的隐藏状态。假设第 t 步的隐藏状态为 $s_t \in R^d$ ，常见方法为将 s_t 和 h_i 点积得到注意力分数（attention score），也称“相似度”或“匹配得分”，如公式 3.1 所示， T 表示转置：

$$e' = [s_t^T h_1, \dots, s_t^T h_N] \in R^N \quad (3.1)$$

随后对当前时间步采用 softmax 方法计算注意力分布 α' ，也称注意力权重（attention weight），该值的大小表明了当前时间步对预测序列的作用大小。如公式 3.2 所示：

$$\alpha' = \text{softmax}(e') \in R^N \quad (3.2)$$

对于得到的注意力权重，采用权重加权求和的方式，结合编码器的隐藏层状态 h_i ，得到需要的上下文向量 c_t ，如公式 3.3 所示， d 表示隐藏层单元数：

$$c_t = \sum_{i=1}^N \alpha_i^t h_i \in R^d \quad (3.3)$$

最后将上下文向量 c_t 和解码器的隐藏层状态 s_t 结合，看成一个 seq2seq 模型，送入解码器预测输出，如公式 3.4 所示：

$$[c_t; s_t] \in R^{2d} \quad (3.4)$$

这里，注意力机制可以给出一个通俗的定义，即可以理解为：对于一组给定的向量集合 values 和一个向量 query，attention 机制可以通过加权求和的方式，依据 query 求取 values。其重点就是如何计算每个“value”的权值，由此形成了不同的 attention 计算方式。

3.2 注意力机制的模型改进

继视觉图像领域提出注意力机制之后，注意力机制在 NLP 领域大放异彩，2014 年到 2015 年这些年，基于 attention 的 RNN 模型大量爆发。到了 2017 年，注意力机制的变种大量出现，其中里程碑之作为 Vaswani^[45]等人发表在 NIPS 上的称为 Transformer 的网络结构，该结构推翻了传统的 Encoder-Decoder 框架，直接构建了只有 attention 机制的网络模型，在机器翻译任务上取得了 state-of-the-art 的成绩。这一小节介绍几种常见的注意力机制改进模型。

3.2.1 在注意力向量的加权求和方式上改进

(1) soft attention

soft attention，又可以称为“全局”attention 或者“动态”attention，是最常见的 attention 种类。以文献^[44]为例，计算注意力分配概率分布的时候，对于源输入句 S 中每一个单词都计算概率，形成概率分布，从而形成上下文向量 c_t ，如 3.1 小节公式 3.3 所示。

(2) hard attention

在 soft attention 普遍流行之后，Xu 等人^[46]又提出了 hard attention。与 soft attention 机制不同，soft attention 是给源输入句中的每个单词都给予一个匹配概率，而 hard attention 是直接从源输入句里找到某个特定单词，将目标句单词和这个单词对齐，而将其他单词的匹配概率硬性地规定为 0。hard attention 一般用在图像中，当图像区域被选中时权重为 1，否则为 0。

(3) local attention

local attention 又称为“半软半硬”attention，对于 soft attention 来说，每次对齐的时候都要考虑编码端之前所有的隐藏状态 h_i ，具有很大的计算开销，因此一种

直接简单的想法就是设置一个大小为 K 的窗口，每次只考虑该窗口内编码端的隐藏输出，归一化为对应的匹配概率，其余部分概率规定为 0，详见文献^[47]。

此外，在注意力向量的加权求和方式上改进的注意力模型还有静态 attention、强制前向 attention 等，详见文献^[48-49]。

3.2.2 在匹配度的计算方式上改进

(1) basic dot-product attention

dot-product attention 即点积注意力，就是常见的 attention score 计算方式，如公式 3.5 所示， e_i 表示 attention score， h_i 表示 Encoder 的隐藏层状态， s 为需要匹配的 Decoder 的隐藏层状态， T 表示转置。

$$e_i = s^T h_i \in R \quad (3.5)$$

(2) multiplicative attention

multiplicative attention 即乘法注意力，将常见的点积计算注意力分数的方式换成了乘法，如公式 3.6 所示， W 矩阵是训练得到的参数，维度为 $d_2 \times d_1$ ， d_2 是 Decoder 隐藏状态 s 的输出维度， d_1 是 Encoder 的隐藏状态 h_i 的输出维度。

$$e_i = s^T W h_i \in R \quad (3.6)$$

(3) additive attention

additive attention 即加法注意力，注意力分数计算方式换成了加法，对两种隐状态 s 和 h_i 分别再训练矩阵 W_1 和 W_2 ，使用激活函数之后再乘以一个参数向量 v 得到注意力分数。如公式 3.7 所示， W_1 的维度为 $d_3 \times d_1$ ， W_2 的维度为 $d_3 \times d_2$ ， v 的维度为 $d_3 \times 1$ ， d_1 ， d_2 ， d_3 分别为 Encoder 的隐藏状态 h_i 的输出维度、Decoder 隐藏状态 s 的输出维度和参数向量 v 的维度，均属于超参数。 T 表示转置。

$$e_i = v^T \tanh(W_1 h_i + W_2 s) \in R \quad (3.7)$$

3.2.3 其他特殊的注意力

(1) self attention

一些特殊的 attention，如文献^[50]中提出的 self attention，在很多任务上，如阅读理解、文本继承、自动文本摘要等取得了十分出色的效果。self attention 即在没有其他额外信息的情况下，句子内部使用 self attention 处理自己，从而获取到句子需要关注的信息。这里介绍 self attention 常见的两种计算方式。

一种是以当前隐藏状态去计算和前面隐藏状态的得分，作为当前隐藏单元的注意力分数，如公式 3.8 所示， W 是经过训练得到的参数，维度为 $d \times d$ ， h 维度为 $d \times 1$ ， d 为隐藏状态 h 的输出维度：

$$e_i = h_i^T W h_i \quad (3.8)$$

另一种是直接以当前隐藏状态去计算得分作为当前单元的注意力分数，这是一种更为常见的计算方式。如公式 3.9 和公式 3.10 所示， W_a 是经过训练得到的参数，维度为 $d \times d$ ， h_i 和 w 均为 $d \times 1$ 的向量， b 为偏置，其中 d 为隐藏状态 h_i 的输出维度。

$$e_i = v_a^T \tanh(W_a h_i) \quad (3.9)$$

$$e_i = \tanh(w^T h_i + b) \quad (3.10)$$

(2) key-value attention

key-value attention 即关键值注意力机制，由 Daniluk 等人^[51]首次提出。简单来说，关键值注意力机制就是将 h_i 拆分成 $[key_i; value_i]$ 两部分，使用的时候只对 key 部分计算注意力权重，加权求和时只使用 $value$ 部分。

3.3 注意力机制在场景文字识别领域的应用

注意力机制自从被提出以来，受到了专家学者的广泛关注，在多个领域都有十分成功的应用，这里重点介绍注意力机制在场景文字识别领域的应用。

注意力机制在文字识别领域的典型代表为 Google 公司提出的基于注意力机制的街景文本识别方法^[52]，模型框架图如图 3-3 所示。

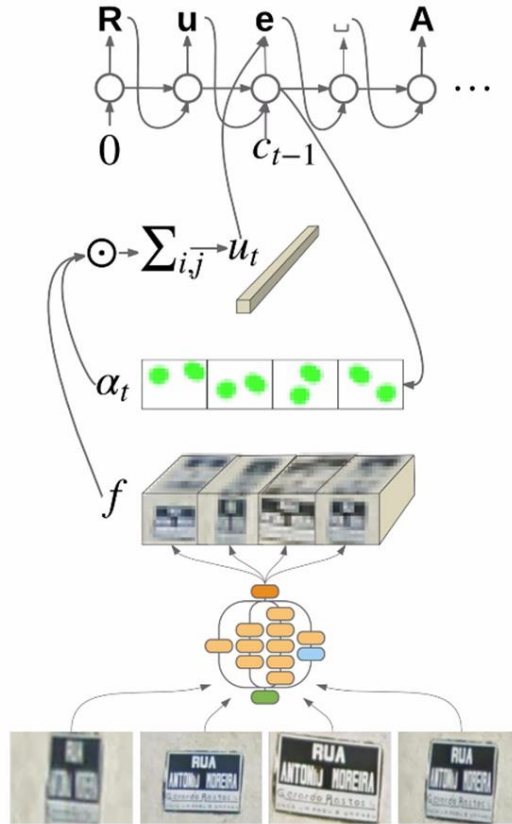


图 3-3 基于注意力机制的街景文本识别

该识别框架是一个 Encoder-Decoder 结构，在 Street View Business Names dataset (SVBN) 和 French Street Name Signs Dataset (FSNS) 两个数据集上取得

了相当好的效果。底层由 CNN 提取原始图像特征以及使用 inception 网络增强特征图构成注意力模型的输入，FSNS 数据集中每张图片有 4 个不同视角，四个视图均通过 CNN 特征提取器，构成特征图 f 。文中使用了空间注意力机制，根据 RNN 神经元的隐藏状态及 CNN 的编码输出向量计算出注意力权重 α_t ，并采用空间加权组合的方式，形成一个固定大小的特征矢量 u_t ，将加权后的特征图送入 RNN 进行解码输出。

Lee 等人^[53]也提出 R^2AM 方法，在无字典约束的情况下，解决自然场景的文字识别问题，整体系统结构如图 3-4 所示。该方法编码阶段采用 recursive CNNs 的方式，recursive 相互作用也可以视为特征间的一种“横向连接”，与普通 CNN 相比增加了深度，同时可以产生更加紧凑的特征响应。解码阶段使用带 Attention Model 的 RNN，直接使用图片进行词汇字符串学习，实现了对无约束自然场景的文字识别。Attention-based 机制使得模型聚焦于输入图片最重要的特征，并且具备更强的可解释性。

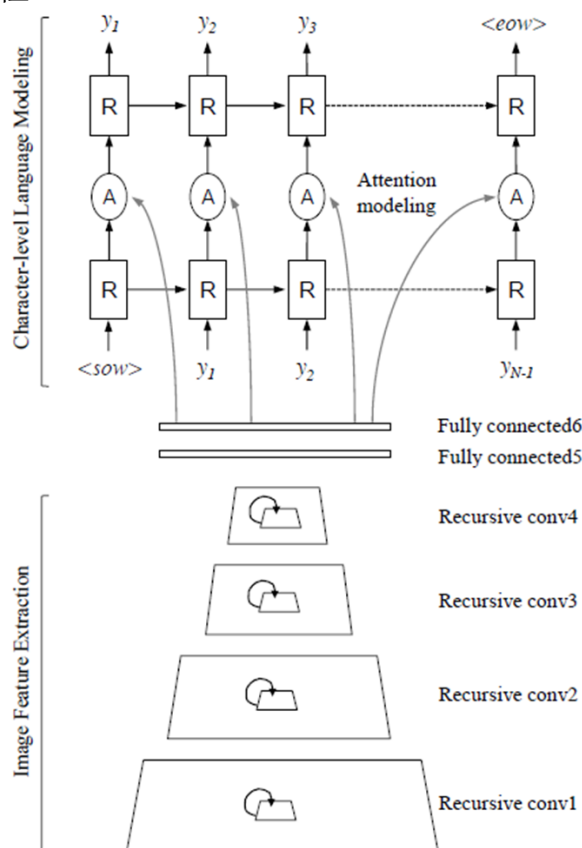


图 3-4 R^2AM 方法

在注意力机制引入场景文字领域并取得一定的成功后，cheng 等人^[36]发现注意力机制在复杂和低质量的图像上表现较差。一个主要的原因是现有方法对于这样的图像不能得到特征区域和目标区域之间的准确对齐，他们称这种现象为“注意力漂移”（Attention Shift）。为了解决这个问题，文章提出 FAN（Focus Attention Network）方法，该方法用聚焦注意力机制来自动拉回漂移的注意力。如图 3-5

所示, FAN 包含两个主要部分: AN 即用于识别字符的普通注意力机制网络, FN 用来评估 AN 的注意力是否与图像中目标区域对齐。在 AN 部分, 产生目标标签和特征间的对齐因子。每一个对齐因子对应一个输入图像的注意力区域。FN 部分, 首先计算每个预测标签的注意力中心, 然后用相应的 *glimpse* 向量预测这个注意力区域, 通过产生在注意力区域可能的分布来使注意力聚焦到目标区域。

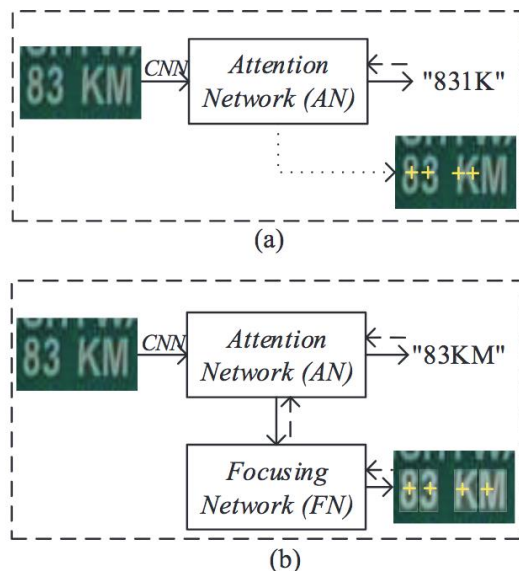


图 3-5 FAN 方法

注意力机制在场景文字识别领域已经得到了成功的应用, 使用注意力机制使得自然场景下的图像文字识别一定程度上有所改善, 但是仍然存在一定的局限, 未来仍有巨大的探索空间。

3.4 本章小结

本章介绍了什么叫注意力机制、注意力机制的变种以及应用场景。注意力机制模仿了人类的视觉选择性处理机制, 经过多年的发展, 注意力机制已经多次进行了模型优化与改进, 以适应不同的任务。主要的改进方式有两种, 分别是在注意力向量的加权求和方式上改进和在匹配度的计算方式上改进。最后介绍了注意力机制在场景文字识别领域的应用。

第四章 基于注意力机制的自然场景文字识别

别

4.1 算法框架描述

麻省理工大学学者 Judd 等人^[54]在研究在一张场景图片中人们到底关注什么这一问题时,发现这样一个现象,在看到同时包括文字和其他事物的图片时,人们的注意力往往会更集中在文字上。对于一张文本图像的识别,需要重点关注文字区域特征的选取。目前所存在的大多数识别算法都采用传统的编解码架构进行上下文特征的选取,但是却忽略了图像空间特征提供的重要信息。本文提出了一种层次化注意力识别网络(HARN, Hierarchy Attention Recognition Network)模型。该模型由基于层次化注意力机制的 Encoder-Decoder 框架构成。在编解码的过程中,不仅通过注意力机制选择了对文本识别来说更重要的图像通道特征,也通过注意力机制更准确地实现了序列上下文特征在原图中的定位。算法整体框架如图 4-1 所示。

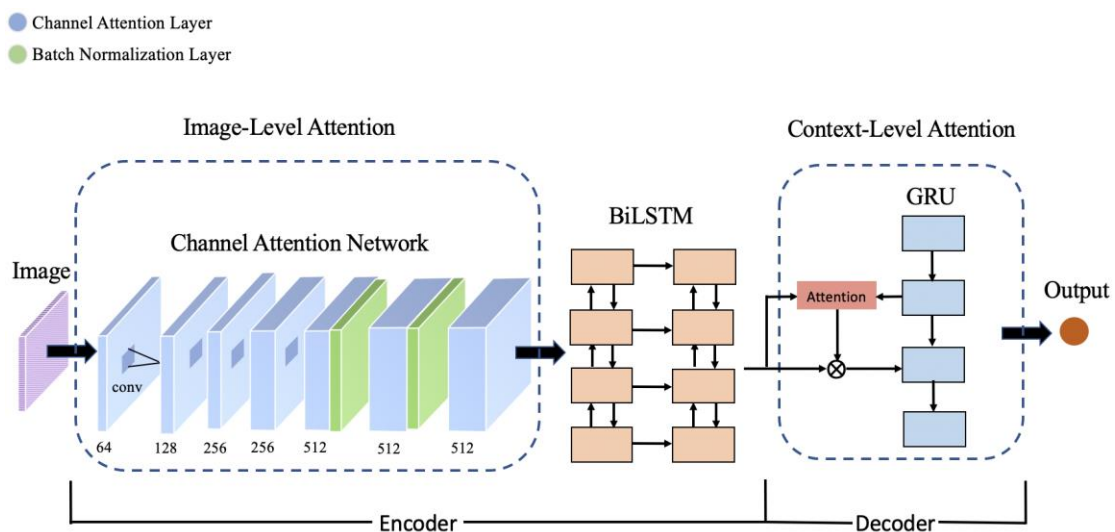


图 4-1 HARN 整体框架

如图所示, Encoder 端中的蓝色方块代表通道注意力层(Channel Attention Layer),绿色方块代表批归一化层(Batch Normalization Layer)。Encoder 采用 Attention-based CNN 和 BiLSTM 构成的二次图像编码方法。编码过程分为两个阶段,第一阶段首先使用 CNN 提取输入图片的局部空间特征,然后通过通道注意力机制(Channel Attention),对提取的局部特征进行特征重标定,选取对识

别字符更重要的特征组合，该特征组合定义为图像通道特征（Image Channel Feature）。由 CNN 和 Channel Attention 共同构成通道注意力网络（CAN, Channel Attention Network），CAN 提供了图片级的注意力选择（Image-Level Attention），该网络更关注于图像自身的空间特征信息，并有选择地强调相互关联的通道图。在编码的第二阶段，将一次编码得到的图像通道特征进一步输入至 BiLSTM 中，BiLSTM 对特征组合进行序列编码，形成序列上下文特征（Sequence Context Feature），这里的序列信息可以看成是一种时序信息。针对图像文本这一特殊对象，通过二次图像编码方法，进行了特征增强，可以更好地捕捉图像自身空间特征以及包含的序列上下文特征。

Decoder 端由 Attention-based RNN 构成，本文使用了门控循环控制单元（GRU），并引入了软注意力机制（Soft Attention），将编码得到的增强特征进行解码预测。由 GRU 和 Soft Attention 共同构成了文本级的注意力选择（Context-Level Attention），即更关注于图像文本中的时序信息，能够更准确地对文本字符与原图中对应位置进行建模。

由 Encoder 和 Decoder 构成的层次化注意力识别网络（HARN），能够有效得通过 Image-Level Attention 以及 Context-Level Attention 分别捕捉图像自身空间信息以及图像文本包含的时序信息，具备更强的特征表达能力，从而提高自然场景下文字识别的准确率。接下来的小节，将会具体介绍各模块设计细节。

4.2 基于 CAN 和 BiLSTM 的图像编码

4.2.1 方法概述

卷积神经网络自提出以来就被广泛应用于图像任务中来，它较全连接网络来说，在提取图像局部特征的同时能够大大减少计算参数量，并且随着网络层数的加深，每层神经元拥有的感受野也会越大。这就带来了感受野越大，在固定视野域中提取到的局部特征容易出现混淆或者歧义的现象。如图 4-2 所示，对于单词“morning”，中间字母“r”和“n”，具有字母“m”的特征，容易被错误识别为字母“m”。但是如果引入上下文信息，在知道前后字母的情况下，就能更好地分辨出这是两个字母。除此之外，上下文信息对于字符串中单个字符的定位也有重要作用。



图 4-2 易造成歧义的局部特征

因此，上下文信息对于自然场景文字识别来说举足轻重。但是，除了关注文字自身带有的上下文信息，图片本身具有的空间局部信息往往容易被忽略。如何能够更好地捕捉图片特征的空间信息并充分利用，是一个值得思考和研究的问题。一般情况下，卷积网络只是简单地提取了空间局部特征，却不能很好地对这些特征进行挑选。这样就会带来不必要的噪声干扰，如会引入多余的背景特征等，影响文字识别的最终准确率。

鉴于以上两点，在本文提出的 HARN 模型的编码阶段，构建了一种基于注意力机制的，带有选择性空间信息和上下文信息的图像编码框架。空间信息由 CAN（Channel-Attention Network）进行选择，CAN 对于提取的图片特征进行特征重标定，选择出更有利于识别的特征信息构成图像通道特征，然后将选择后的特征组合输入到两层的 BiLSTM 网络中去提取上下文信息，形成序列上下文特征，这样就完成了图像编码过程。

4.2.2 基于 CAN 的空间局部特征提取

受^[55-56]的启发，本文中 Attention-based CNN 使用了通道注意力选择（Channel Attention），即更关注局部空间特征图中哪些特征更重要。CNN 的每一层输出都是一个 $C \times H \times W$ 的特征图， C 是卷积核的数量，同时也代表通道数和特征数。 H 和 W 分别代表该层输出特征图的高和宽。对于尺寸为 $H \times W$ 的二维平面，Channel Attention 在所有的平面维度上权重相同，在 channel 维度上学习不同的权重。通道注意力选择可以理解为在局部空间特征中选择图片的不同特征。

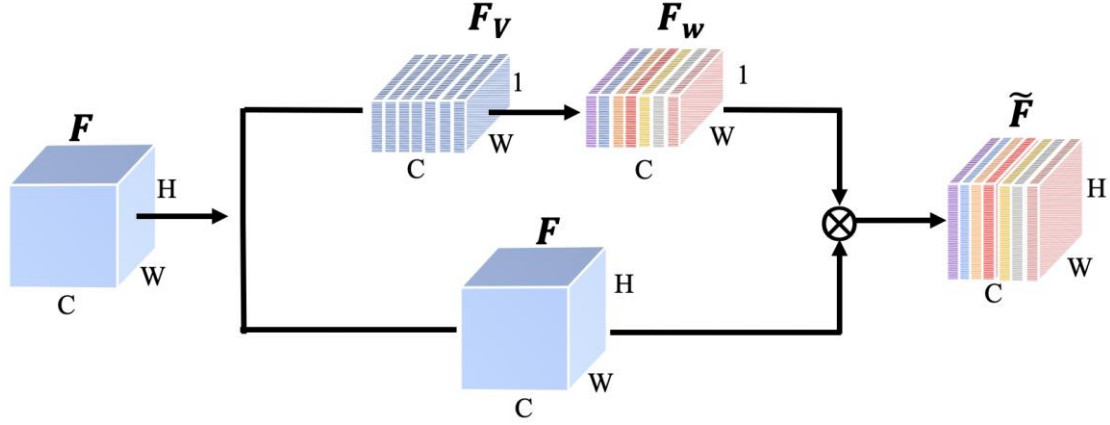


图 4-3 Channel Attention Module

如图 4-3 所示，为一个通道注意力选择模块（Channel Attention Module）。对于一个 $C \times H \times W$ 的卷积结构 F ，在 H 维度上进行压缩，得到输出为 $1 \times W \times C$ 的向量 F_V ，这里采用的是最简单的加和求平均算法，如公式 4.1 所示：

$$F_V = \frac{1}{H} \sum_{i=1}^H F(i) \quad (4.1)$$

F_V 可以看成是 C 个 $1 \times W$ 的向量组成，并定义为 $X=[x_1, x_2, \dots, x_C]$ 。在 x_1 到 x_C 之间，我们通过定义一个 Channel Attention 算法进行注意力权重的计算，受^[57]启发，我们采用其中的 self-attention 计算方式，即将每个 $1 \times W$ 的向量通过两层的全连接层来获得权重，以此来减少参数量。Channel Attention 算法如算法 4-1 所示，其中 W_1 的维度为 $2W \times W$ ， W_2 的维度为 $W \times 2W$ ， b 的维度为 $2W \times 1$ ， T 表示转置。

算法 4-1 Channel Attention 算法

输入： F_V 向量组合 $X=[x_1, x_2, \dots, x_C]$

输出： F_w 权重组合

```

1: def Channel Attention ( $x_i$ )
2:   for  $x_i \in [x_1, x_2, \dots, x_C]$  do
3:     得分  $score \leftarrow W_2 \tanh(W_1 x_i^T + b)$ 
4:     得分集合  $scores[] \leftarrow$  得分  $score$ 
5:   end for
6:    $F_w$  权重组合  $\leftarrow$  softmax 归一化得分集合  $scores[]$ 
7: return  $F_w$  权重组合

```

通过算法 4-1 获得权重组合 F_w 之后，图 4-3 中 F_w 的不同色块代表不同的权重值，将权重值作为乘数因子乘到卷积结构 F 的每个通道上，作为下一层的输入 \tilde{F} 。通过权重的不同，对不同的特征进行选择，增强重要特征，削弱不重要特征，从而让提取的特征具有更强的指向性，选择出更有利于文本识别的特征组合。

Channel Attention 模块可以布置在卷积层中, 由 Channel Attention Module 和多层 CNN 共同构成了 Channel Attention Network (CAN)。因为该网络更关注于图像本身被提取到的空间通道特征, 因此定义为 Image-Level Attention。

本文中 CNN 的结构设计参考了 Shi^[58]等人的 CRNN 模型, 默认使用 7 层的卷积网络, 为了方便对比实验的进行, 本文的模型结构在设计上尽量与其保持一致。CNN 各层详细参数如表 4-1 所示。其中, maps 代表输出通道数, k 代表卷积核尺寸大小, s 代表步幅, p 代表 padding 大小。BN 代表 batch normalization。整个 CAN 网络的输入为一张 $3 \times 32 \times 280$ 的图片, 最终输出为一个 $512 \times 1 \times 71$ 的特征图, 至此完成第一阶段编码。CAN 的输出结果将作为第二阶段编码的输入, 送入 BiLSTM 中。

表 4-1 CNN 各层详细参数

层类型	参数	尺寸(C×H×W)
Input	-	$3 \times 32 \times 280$
Convolution	maps:64, k: 3×3 , s:1, p:1	$64 \times 32 \times 280$
MaxPooling	k: 2×2 , s:2	$64 \times 16 \times 140$
Convolution	maps:128, k: 3×3 , s:1, p:1	$128 \times 16 \times 140$
MaxPooling	k: 2×2 , s:2	$128 \times 8 \times 70$
Convolution	maps:256, k: 3×3 , s:1, p:1	$256 \times 8 \times 70$
Convolution	maps:256, k: 3×3 , s:1, p:1	$256 \times 8 \times 70$
MaxPooling	k: 2×2 , s: 2×1 , p: 0×1	$256 \times 4 \times 71$
Convolution	maps:512, k: 3×3 , s:1, p:1	$512 \times 4 \times 71$
BN	-	$512 \times 4 \times 71$
Convolution	maps:512, k: 3×3 , s:1, p:1	$512 \times 4 \times 71$
BN	-	$512 \times 4 \times 71$
MaxPooling	k: 2×2 , s: 2×1 , p: 0×1	$512 \times 2 \times 72$
Convolution	maps:512, k: 2×2 , s:1, p:0	$512 \times 1 \times 71$
BiLSTM	hidden unit:256	$256 \times 1 \times 71$
BiLSTM	hidden unit:256	$256 \times 1 \times 71$

4.2.3 基于 BiLSTM 的序列上下文特征提取

对于单张图片, CAN 的输出为一个 $512 \times 1 \times 71$ 维的特征图。将其分解为 71 个 512 维的向量, 因为原始图片的高度为 32, 缩小了 4 倍, 所以可以理解每个向量对应原始图像一块 32×32 的区域, 也就是其对应窗口图像的一个特征表示,

如图 4-4 所示。经过两层的 BiLSTM 之后，输出特征为 $256 \times 1 \times 71$ ，其中 256 是 BiLSTM 隐藏节点的个数。

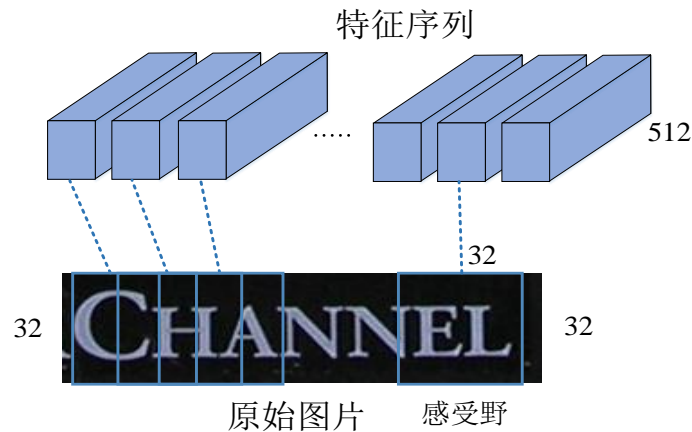


图 4-4 CAN 特征向量与原始图像对应示例

在 CAN 对原始图像进行第一阶段编码之后，使用两层的 BiLSTM 对 CAN 输出得到的图像通道特征进行第二阶段编码，以获取每个窗口的序列特征。BiLSTM 是由两层方向不同的 LSTM 构成的结构，输入为 CAN 输出得到的图像通道特征，输出同样长度的向量序列，称为序列上下文特征，这里可以看成是通过 BiLSTM 进行序列特征的增强。对于字符序列的预测，如果能够同时捕获前面若干输入和后面若干输入的信息，结果将更加准确，所以在这里使用双向的 LSTM。对于 BiLSTM，在 Forward Layer 从 1 时刻到 t 时刻进行正向计算并保存每个时刻前向隐藏层输出。对应的，在 Backward Layer 计算从 t 时刻到 1 时刻的反向，也保存每个时刻反向隐藏层的输出。最后在每个时刻将 Forward Layer 和 Backward Layer 的对应位置的输出进行拼接，得到最终的输出序列。一个 BiLSTM 的示意如图 4-5 所示。

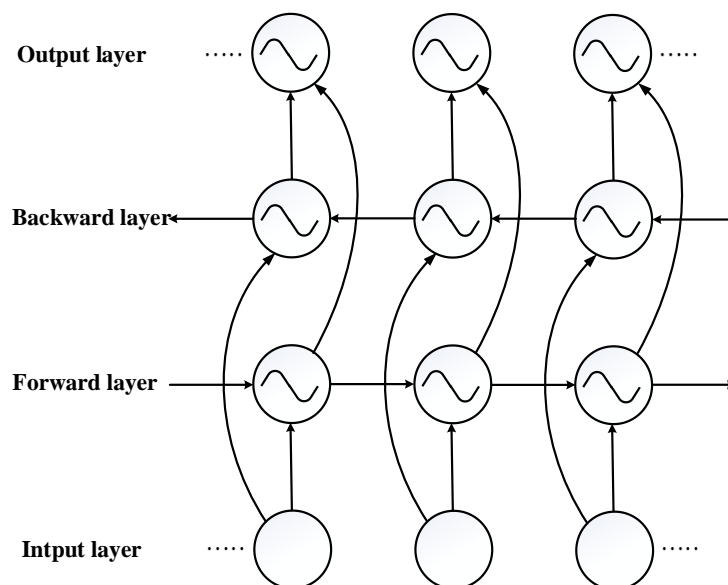


图 4-5 BiLSTM 结构示意图

4.3 基于 ASGN 的图像解码

4.3.1 方法概述

解码器由 Attention-based RNN 构成, 本文构建了一种基于注意力的序列生成网络 (ASGN, Attention-based Sequence Generation Network), 同时完成字符定位与识别。在这里, 该注意力机制定义为 Context-Level Attention, 即更关注于图像中包含的时序文本信息, 对文本字符与原图中对应位置进行建模。

ASGN 通过计算当前输出与输入之间的相关度, 显示地建模网络的当前关注点。由于当前输入为二次编码之后得到序列上下文特征, 每个序列上下文特征与原始图像的某块区域一一对应, 所以计算出当前的预测字符与每个序列上下文特征之间的相关性之后, 就可以大致得到字符在原始图像中的位置。对比 CRNN 模型最后的转录层, 即 CTC 模型^[59], ASGN 有以下优点:

(1) ASGN 引入了注意力机制, 能够显示地对输入和输出进行位置关系的建模, 这一步对于字符定位来说非常重要, 可以更直观地解释当前状态下网络模型的关注点。

(2) CTC 的一个重要特征是多对一的输入和输出关系, 然而这种设定并不理想。多对一的情况意味着输出的时间步长要比输入短, 对于文字识别来说, 可能会出现一个上下文向量中包含多个字符, 导致预测的字符数量大于输入向量的情况, 这时候 CTC 就不奏效了。

(3) 由于 CTC 通常采用集束搜索来搜索可能的预测字符序列, 所以占用的计算资源较多, 预测速度会慢。解码器 ASGN 引入注意力机制, 去掉了 CTC 模型, 模型小, 训练速度快。

编码器采用的是门控循环单元 (GRU), 注意力单元利用 GRU 的隐藏层状态, 查询 (query) 与当前目标相关的编码向量。GRU 利用这些向量合成上下文向量 (context vector), 同时使用上下文向量和上一次的分类结果更新自身隐藏层状态, 并生成当前时刻的预测输出。重复执行以上过程, 直到输出结束符时停止, 所得到的序列即为识别出的字符序列。

4.3.2 ASGN 网络设计

ASGN 使用一种带注意力机制的循环网络 (Attention-based RNN) 对编码得到的序列特征进行解码, 本文中使用了单层的 GRU 网络, 并引入 Soft Attention 机制, 直接生成目标序列 (y_1, y_2, \dots, y_N)。基于注意力机制的 ASGN 预测输出过程如图 4-6 所示。

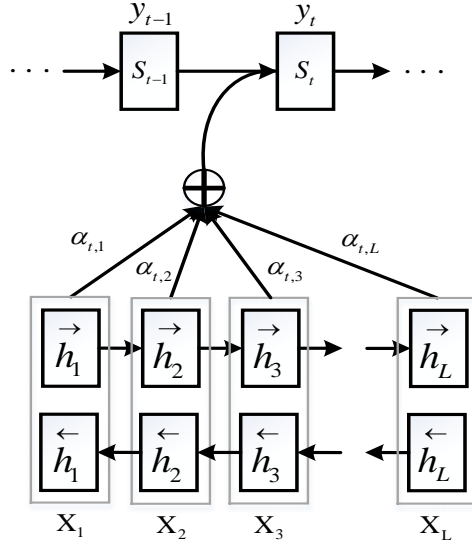


图 4-6 基于 attention 的 ASGN 预测输出过程

首先根据一定的规则求取 attention 的权重。在每一个时间步 t ，都要给注意力打分以获取权重，如公式 4.2 所示，注意力分数定义为 $e_{t,i}$ ， h_i 表示由编码器得到的隐藏状态， s_{t-1} 表示解码器上一步预测输出的隐藏状态， b 表示偏置，与 W_s 和 W_h 一样都是可训练的参数矩阵。

$$e_{t,i} = \tanh(W_s s_{t-1} + W_h h_i + b) \quad (4.2)$$

得到注意力分数之后，对其进行归一化，得到当前关注的权重向量 $\alpha_{t,i}$ ，如公式 4.3 所示， L 为编码得到的序列上下文特征向量的长度：

$$\alpha_{t,i} = \exp(e_{t,i}) / \sum_{j=1}^L (\exp(e_{t,j})) \quad (4.3)$$

接下来再将注意力权重 $\alpha_{t,i}$ 乘以编码器输出的特征矩阵，实质就是将 71 个字符，根据注意力权重合并成 1 个最大概率的字符。如公式 4.4 所示，得到 glimpse 向量 g_t ，也表示上下文向量：

$$g_t = \sum_{i=1}^L (\alpha_{t,i} h_i) \quad (4.4)$$

这里我们做了一个词嵌入 (embedding) 进行升维，如公式 4.5 所示， y_{prev} 表示上一步的输出 y_{t-1} 的嵌入向量。训练阶段的时候，使用上一步的 groundtruth 作为词嵌入，如果是测试阶段，则直接使用上一步的预测输出作为词嵌入。

$$y_{prev} = \text{Embedding}(y_{t-1}) \quad (4.5)$$

联合上下文向量 g_t ，词嵌入向量 y_{prev} 和 GRU 上一步的隐藏状态 s_{t-1} ，我们使用 GRU 更新 s_t ， s_t 表示在当前时间步 t 时刻 GRU 的隐藏状态。计算如公式 4.6 所示：

$$s_t = \text{GRU}(y_{prev}, g_t, s_{t-1}) \quad (4.6)$$

设解码器生成的最大时间步数为 T ，在预测到序列结束标记“EOS”时就停止预测。对于每一个时间步 t ，输出 y_t 定义为公式 4.7：

$$y_t = \text{softmax}(W_{out} s_t + b_{out}) \quad (4.7)$$

这里, W_{out}, b_{out} 也是可训练的参数。最后形成的目标序列为 (y_1, y_2, \dots, y_N) 。

4.4 系统损失函数设计

本文提出的基于层次化注意力机制网络 HARN 是一个端到端的联合训练模型。对于解码器 ASGN, 训练集表示为 $D = \{y_i, \hat{y}_{i,t}\}$, $i = 1, \dots, N$ 。 I 表示单张图片, \hat{y} 表示该张图片对应的 ground truth。因此我们的目标函数是将 D 的条件概率的负对数似然最小化, 表示如公式 4.8 所示:

$$L = - \sum_{i=1}^N \sum_{t=1}^{|\hat{y}_i|} \log p(\hat{y}_{i,t} | I_i; \theta) \quad (4.8)$$

其中, $\hat{y}_{i,t}$ 是 I_i 中第 t 个字符的 ground truth, θ 是包含模型中所有参数的向量。ASGN 的输出一共有 37 类, 包括 26 个英文字母, 10 个数字和一个表示结束标志的“EOS”标志符。

4.5 本章小结

本章对本文提出的方法做了详细介绍, 包括图像的编码和图像解码两部分, 在详细描述了整个算法框架之后, 也阐述了图像编码和解码的网络结构设计。图像编码主要分为两个阶段, 第一阶段由基于 Channel Attention 的 CAN 对输入图像编码, 第二阶段由两层的 BiLSTM 进一步进行特征加强。在图像解码部分, 对基于 Soft Attention 的解码器的原理和网络结构做了详细介绍, 包括字符的定位以及识别过程。

第五章 实验与分析

5.1 实验使用数据集介绍

5.1.1 训练集

为了尽量与 CRNN 实验配置保持一致，所以本文实验中也采用 Jaderberg 等人发布的合成数据集 Synth90k^[29]当作训练集。该数据集使用图像合成引擎模拟自然条件下的遮挡、阴影、边框等真实场景图片，生成一个人工合成数据集。人工合成的优势就在于，每张数据不需要人工标注就可以准确的知晓 label 的位置和信息，这为自然场景下的文本识别提供了便捷的数据获取方式。Wang 等人^[60]和 Jaderberg 等人^[29]也都采用人工合成的数据去训练 CNN 的文本识别网络，并取得了不错的效果。Synth90k 数据集包含 900 万张图像，涵盖 90k 个英文单词，其中 720 万张被划分成训练集，90 万张用做验证集，剩余 90 万张用做测试集。如图 5-1 所示，为数据集中人工合成数据实例。

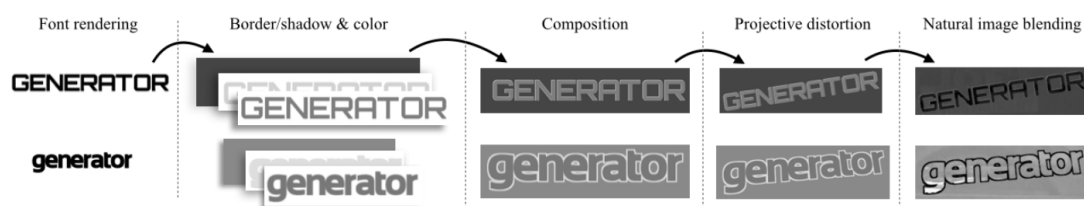


图 5-1 人工合成数据实例

5.1.2 测试集

在本实验中，除了使用 Synth90k 数据集中的测试集，我们还采用一些公开的基准数据集充当性能评估的测试集，主要使用了 4 个数据集，分别是 ICDAR2003^[37]、ICDAR2013^[39]、IIIT5K^[22]以及 SVT^[11]。其中 ICDAR2003、ICDAR2013 和 IIIT5K 数据集已经在本文 2.5 小结具体介绍过了，在此不再赘述。SVT 数据集的图像来源于谷歌街景，一共包含 350 个高分辨率图像，平均大小为 1260×860。其中 100 张图片用于训练，250 张图片用于测试，并且只提供单词级别的边界框以及不区分大小写的文本字符串标注。

5.2 图像预处理

一张图片在输入网络之前需要进行一些处理才能更好的适应网络模型的需要，本文也对训练集图片和测试集图片做了简单的处理。首先，我们没有将图片

转换为灰度图，因为想要尽可能地保存更多图片的信息，转化为灰度图像虽然减少了一定的计算量，但是却容易丢失一部分图像特征信息，因此输入网络模型的图片仍然保留为 3 通道的 RGB 图像。

其次，对于训练集图片进行了缩放操作，图像缩放是为了在模型中保持固定的高度，使用双线性插值算法将输入图片统一缩放到 32pix 高，考虑到神经网络的批量训练，所以将图片宽度进行统一。对于长度大于 280pix 的图片，统一规定为 280pix，长度不足的加上 padding 之后达到 280pix，padding 部分的像素取图片每个通道像素值的均值，最后效果如图 5-2 所示。

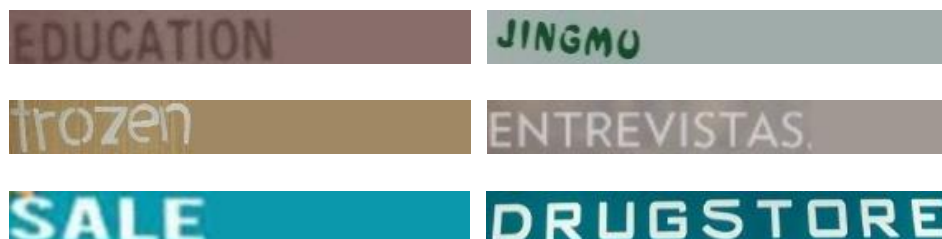


图 5-2 训练集图片预处理

经编码器编码后输出长度为 71 的特征矩阵，完全可以满足绝大部分中英文字符串的长度。对于测试图片，图片的高度仍然缩放到 32pix，但是宽度不做限制。

5.3 实验设置

5.3.1 实验环境

本文是在 PyTorch 框架下完成整个模型的搭建，实验中使用的基礎环境是 CUDA 8.0 以及 CuDNN v7，因此模型使用的是 GPU 加速，其型号为 NVIDIA TITAN XP GPU。CPU 为具有 64 位 40 核的 2.20GHz Intel(R) Xeon(R) CPU E5-2630 v4。在训练期间，所有图像都缩放到 32×280 用以加速训练过程。具体实验环境如表 5-1 所示。

表 5-1 实验环境

CPU	Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz
GPU	NVIDIA TITAN XP GPU
显存	12G
操作系统	Ubuntu 16.4
CUDA	8.0
Cudnn	7.1.2
开发语言	Python 3.6
深度学习框架	Pytorch 0.4.0
其他环境依赖库	OpenCV PIL (Pillow) TorchVision matplotlib

5.3.2 训练策略

(1) 学习率优化

对于模型训练来说，学习率是一个非常重要的超参。一般情况下在训练轮数开始时学习率可以设置大些以保证模型以较快的速度收敛，也避免过早陷入局部最优，在训练的后期可以降低学习率进行微调，以保证收敛过程能达到最佳状态。实验中采用了指数减缓（exponential decay）的学习率调整方式，具体计算公式为 5.1 所示：

$$lr = base_lr \times 0.5^{(\text{epoch}/\text{step})} \quad (5.1)$$

其中，初始学习率 $base_lr$ 为 0.001，epoch 为当前 epoch 的迭代轮数，step 为 10。

同时，实验中采用的优化器为 Adam，它和传统的随机梯度下降算法不同，不再保持单一的学习率不变，而是利用梯度信息设计不同的自适应学习率，从而能够在训练时通过数据不断地迭代来更新神经网络的权值。实验中设置 α 为 0.001， β_1 为 0.5， β_2 在稀疏梯度中应该设置为接近于 1 的数，因此设置为 0.999。

(2) 数据增强

数据增强（Data Augmentation）是一种神经网络训练中常见的防止过拟合的手段。它通过对数据集加噪声、旋转、缩放等操作实现数据集的扩充，扩充的数据集能够拥有更多丰富的新数据，帮助减少训练和测试误差，增强神经网络的鲁棒性，使其具有更强的泛化性能。实验中使用了 `imgaug`^[61] 数据增强库，每次使

用 1 到 3 个增强方式来处理图片，每个 batch 中增强方式的顺序是随机的。经过多次尝试，最后挑选出几种比较合适的数据增强方式，如表 5-2 所示：

表 5-2 数据增强方式选取

Invert	将每个像素值 p 变成 $255-p$
Add	每个像素点随机加上一个值
Multiply	每个像素点乘以一个值
Dropout	随机去掉图像中的一些像素点
GaussianBlur	高斯扰动
AverageBlur	从最近邻像素中取均值扰动
MedianBlur	通过最近邻中位数来扰动
Emboss	浮雕效果
EdgeDetect	边缘检测

(3) Dropout

在训练阶段，实验中使用了 Dropout 作为一种防止过拟合的手段。Dropout 可以通俗理解为在前向传播的时候，设置一定的概率，使某个神经元停止工作，它主要针对训练阶段的全联接层，随机去除掉某些连接以增强模型的泛化性能。在每个 batch，随机忽略一部分的隐藏节点，即设置隐层节点值为 0，可以减少过拟合现象。

在实际实验中，训练阶段的 Dropout 设置为 0.5，测试阶段没有使用 Dropout。但是由于本文大量使用了卷积神经网络和 ReLu 函数，卷积神经网络有自身的稀疏性，只包含少量全联接层，所以尝试了此方法后效果提升并不明显。

5.4 实验结果

5.4.1 消融实验

本文选择了 Shi^[58]等人提出的 CRNN 模型作为对比模型，该模型作为一个经典模型，在主流文字识别算法当中，尤其是无约束文字识别领域取得了 state-of-the-art 的水平，因此具有一定的可比性。实验过程中，在 Synth90 训练集中随机挑选了 100 万张图片作为训练集，1 万张作为验证集，1 万张作为测试集。为了公平起见，用 100 万张数据集同时对 CRNN 和本文提出的模型 HARN 进行训练，并进行测试。对 CRNN 模型的训练大约为 38 个小时，HARN 模型的训练大约为 40 个小时，从这一点可以看出虽然 HARN 在模型结构上比 CRNN 稍微复杂，但是模型训练完全收敛的时间相差不多。

为了验证本文算法的优越性，我们进行了消融实验。如表 5-3 所示，CRNN 作为本次实验的基准实验（baseline experiment），不做任何结构上的改变；Attention-CNN 表示只在编码阶段的 CNN 上加入本文算法提出的 channel attention 机制，构成 CAN 网络结构；Attention-RNN 表示只在解码阶段的 RNN 上加入本文算法提出的注意力机制，构成 ASGN 解码器；HARN 即表示本文提出的完整的层次化注意力机制模型。表 5-3 列出了消融实验的结果，评价标准使用了单词准确率即全匹配准确率，以及平均编辑距离。

表 5-3 消融实验

方法	单词准确率	平均编辑距离
CRNN	71.92	0.1249
Attention-CNN	72.34	0.1220
Attention-RNN	72.10	0.1231
HARN	72.68	0.1192

从实验结果可以看出，相比于 CRNN，不论是在编码阶段还是解码阶段引入注意力机制，均能取得更好的识别性能，尤其是在 CNN 上加入本文提出的 channel attention 算法，比只在 RNN 上加入注意力机制高出了 0.24% 的精度，比单纯的 CRNN 模型高出了 0.42% 的精度。当在编码和解码阶段同时使用层次化注意力机制后，取得了最高的单词准确率，比单纯的 CRNN 模型高出了 0.76% 的精度。从平均编辑距离来看，HARN 也取得了最小的编辑距离，这意味着更佳的识别性能。

5.4.2 模型优化实验

本文提出的 HARN 算法是一个灵活的算法模型，为了进一步探讨该算法与神经网络层数之间的关系，我们对现有模型进行了改造以寻找最佳的网络层数配置。同时为了防止神经网络层数的加深而导致梯度消失或者梯度爆炸，我们使用了 ResNet 网络^[62]。如表 5-4 所示，同样使用 CRNN 作为本次实验的基准实验，不做任何结构上的改变；HARN-base 表示 CNN 部分只使用了 4.2.2 小节中定义的七层卷积神经网络；HARN-18 表示将七层卷积神经网络替换成 ResNet-18 网络，HARN-34 表示将七层卷积神经网络替换成 ResNet-34 网络，以此类推。模型优化实验结果如表 5-4 所示。

表 5-4 模型优化实验

方法	单词准确率	平均编辑距离
CRNN	71.92	0.1249
HARN-base	72.68	0.1220
HARN-18	72.94	0.1205
HARN-34	72.98	0.1202
HARN-50	73.70	0.1149
HARN-101	72.60	0.1226

如表中所示，相比于 CRNN，使用简单七层卷积神经网络的 HARN 模型在识别精度上有较大幅度提升。随着网络层数的加深，HAN-18 模型表现出良好的性能，然而 HARN-34 在 HARN-18 的基础上基本没有太大提升。表现最好的模型为 HARN-50，比 HARN-base 高出了 1.02% 的精度，比 CRNN 高出了 1.78% 的精度。但是随着卷积层数进一步加深 101 层，由于层数过深，HARN-101 模型整个训练过程十分缓慢，并且模型结构太过复杂导致识别精度反而下降。

5.4.3 标准数据集实验

为了验证本文算法在自然场景下文本识别的优越性，我们在一些业内公开的标准数据集上进行了测试。实验是在无字典约束的模式下进行的，实验中没有使用任何字符级的文本标注，也没有对特定数据集进行 fine-tuning，只使用了 Symh90k 人工合成数据集。但是实验证明，虽然只使用了合成数据训练模型，HARN 仍然能够对真实场景的数据表现出良好的识别性能。整个模型的收敛一共花费了大约 3 天的时间，只通过一次在合成数据集上的训练，在包括 IIIT5K，SVT，IC03 和 IC13 数据集在内的标准基准数据集上进行测试。表 5-5 总结了本文方法 HARN 在这四个标准数据集上的识别结果，并将结果与部分同期较好的模型方法进行了比较。这里使用了 HARN-base 基本模型来进行比对，即 CNN 部分只使用了七层卷积神经网络。

表 5-5 标准数据集实验

方法	IIIT5K	SVT	IC03	IC13
Bissacco et al. ^[19]	-	78.0	-	87.6
Lee and Osindero ^[63]	78.4	80.7	88.7	90.0
Yin et al. ^[64]	78.2	72.5	81.1	81.4
Jaderberg et al. ^[30]	-	80.7	93.1	90.8
Jaderberg et al. ^[65]	-	71.7	89.6	81.8
Shi, Bai, and Yao ^[58]	78.2	80.8	89.4	86.7
Ours	78.4	81.4	89.9	87.0

由实验结果可以看出,相比于表中列出的其他模型,HARN 在 IIIT5k 和 SVT 两个数据集上都取得了最好的识别精度,而 IC03 和 IC13 两个数据集上取得最好成绩的是 Jaderberg 等人^[30]提出的模型,但是该模型只能识别 9 万个单词,无法预测词汇表外单词,严格来说并不是真正意义上的无字典约束模式。相比于 Bissacco 等人^[19]提出的 PhotoOCR 系统,该模型使用了带有字符级标注的 800 万训练数据,而 HARN 只在包含单词级标注的人工合成数据集上进行了一次训练,就取得了不错的识别性能,这也说明了 HARN 是高性能且低成本的算法。同时从表中也可看出,HARN 在四个数据集上的表现均优于 CRNN 模型。SVT 和 IC03 数据集中的文本大多为水平方向布局,HARN 的识别精度较 CRNN 模型提升了 0.5% 以上,但是由于 IIIT5K 和 IC13 两个数据集包含较多不规则、多朝向文本,因此精度较 CRNN 模型并没有太大程度的提升,如何进一步增强 HARN 的通用性,尤其是应对复杂场景下的不规则文本识别,这也是 HARN 模型接下来需要提高和改进的方向。

5.5 实验分析

5.5.1 识别过程可视化

本文提出的算法是基于注意力机制的层次化模型,其最大优点就是可以显示地观察到当前时刻神经网络所关注的位置。为了能够更直观地体现本算法的优势,展示算法的性能,我们选取了一部分测试样例,对 ASGN 进行了可视化显示,帮助进一步理解本文算法的优势。

如图 5-3 所示,为识别正确的测试样例。选取的测试样例都是带有一定挑战性的文本,(a)、(b)、(c) 三个样例是典型的例子,图 5-3 (a) 代表弯曲文本,图 5-3 (b) 包含具有阴影的文本,图 5-3 (c) 代表字符和背景区分度不明显的文本。每张图都有其对应注意力机制输出结果的可视化显示,不同亮度的区

域表示当前时刻注意力机制的关注区域，即代表此时最可能出现的字符，亮度高低表示此区域注意力的强弱，亮度越高表示该区域字符出现的概率越大。由图可以看出，不论字符布局是一定程度弯曲，还是水平布局，算法模型都能准确地预测识别。对于自然场景的真实图片，或者是人工合成图片，也具有较强的识别性能。算法的注意力机制总能较为准确地关注到应该关注的区域。

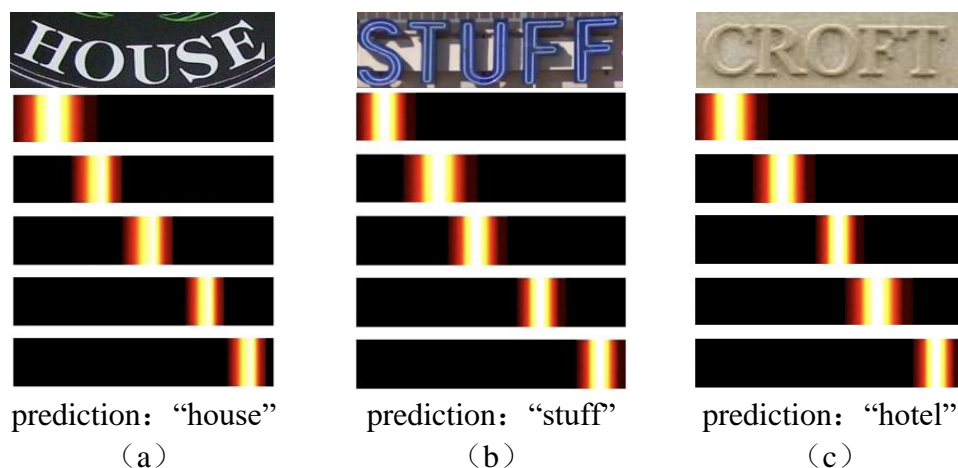


图 5-3 识别正确的样例

然而，虽然本算法的识别性能具有较强的鲁棒性，但是在识别过程中也有一些较为常见的错误。如图 5-4 所示，列举了一些识别错误样例。图 5-4 (a) 和图 5-4 (b) 类似于手写体和花体字符，具有较多的连笔，字符间特征区别度较小，所以图 5-4 (a) 中模型将字母“a”识别成“o”，图 5-4 (b) 中将字母“tt”识别成“ll”。图 5-4 (c) 图片的分辨率较低，而且也存在连笔现象，导致模型定位出现偏差，字符个数识别错误，将“venus”识别成了“venuus”，多识别出单个字符“u”。不过，通过识别结果可视化可以看出，字符的定位无太大偏差，注意力机制基本能够准确定位到当前关注区域。这对于后期位置的矫正提供了便利。

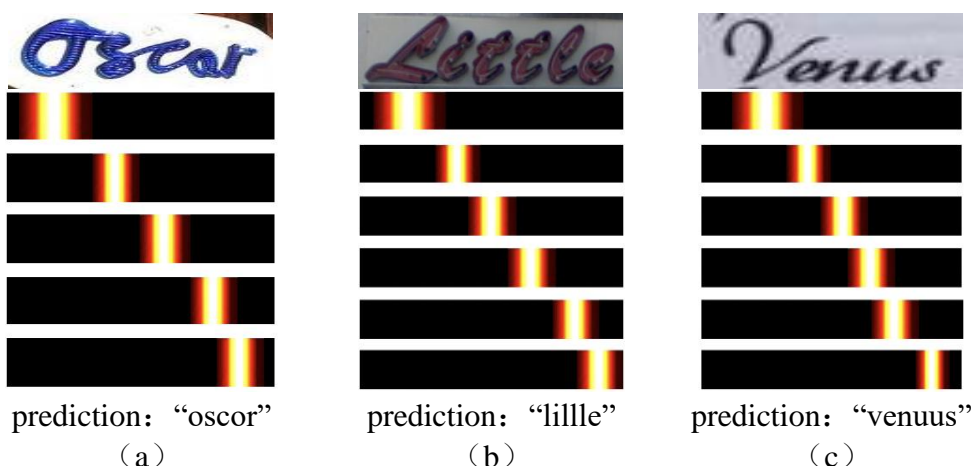


图 5-4 识别错误的样例

5.5.2 泛化能力分析

通过以上实验可知，本文方法在人造数据以及自然场景的标准数据集上都能取得较好的水平，为了验证该算法在复杂场景下的迁移能力，搜寻了一些特殊文本图片，如不规则文本、艺术字、带水印文字等，该数据集测试图片为 300 张。实验中同时对比本文 HARN 模型与 CRNN 模型。CRNN 的识别准确为 45.78%，本文 HARN 模型的识别准确率可以达到 46.89%，高出 CRNN 模型 1 个点的精度。

识别结果如图 5-5 所示。可以看出，针对这些极端场景的文本，两个模型的识别能力都存在一定的欠缺，识别精度不高，过于复杂的文本难以识别准确。常见的问题为易混淆字母识别错误，比如，由于一些艺术字的写法，字母“p”会识别成字母“d”，字母“c”会识别成字母“o”等。模型也无法区别前景与背景，如“viva”图片中由于数字“88”上半部分的出现，模型错误识别为“vova”。但是从识别结果可以看出，HARN 的识别情况明显优于 CRNN 模型，对于复杂情况下的文本识别具有更强的泛化能力和抗干扰能力。















 “mcdull”	 “llidule”
 “yclym6688”	 “ay”
 “r”	 “viva”
 “rong”	
Our model	
 “medull”	 “blidle”
 “yolym6688”	 “ar”
 “3”	 “vova”
 “rm”	
CRNN model	

图 5-5 复杂文本的识别效果

5.5.3 综合性能分析

为了进一步体现本文算法所具优势，将从以下 4 个方面进行综合比较，分别为是否可端到端训练，是否有字符级标注，是否有字典约束，模型参数量大小。

（1）端到端训练

对于传统文字识别算法来说，往往需要单独训练多个子模块才能完成图像到字符的识别。本文提出的方法基于深度神经网络，同 CRNN 一样，具备可端到端训练的能力，不需要任何前处理或者预训练过程。这对于模型训练来说提供了极大的便利，模型更易训练，移植也更加方便。

（2）是否有字符级标注

很多模型的训练不仅需要单词级别的人工标注，还需要字符级别的标注，这无疑需要耗费巨大的人力，工作也较为繁琐。但是本文提出的算法只需要提供单词级别的标注即可，免去了复杂的字符级标注。

（3）无字典约束

这项指标用于评价模型是否受限于特定的字典，而不能处理字典以外的文本序列。本算法在这一方面具有很大的自由度，它不受限于识别的输出空间，可以识别任意的字符序列。

（4）模型参数量

模型参数量对于衡量一个模型的好坏程度来说至关重要，通常情况下，我们希望参数量越少越好，这意味着模型所需的存储空间越小。CRNN 模型基于卷积神经网络，没有全连接层，卷积层权值共享，循环神经网络的参数也只是一个循环元的参数，因此一共有 830 万个参数。而本文提出的模型在 Decoder 端多添加了 GRU 单元，GRU 自身包含全连接结构，因此参数量有少许增加。但对于识别性能的提高来说，是值得的。对于 HARN-base 模型来说，参数量大约为 840 万。

表 5-6 详细地展示了本文提出的方法与各主流方法之间以上各指标的比较，M 代表百万。

表 5-6 不同方法之间各指标比较

模型名称	端到端训练	字符级标注	无字典约束	模型参数量
Mishra et al. ^[21]	✗	✗	✗	-
Wang et al. ^[60]	✗	✗	✓	-
Goel et al. ^[24]	✗	✓	✗	-
Yao et al. ^[66]	✗	✗	✓	-
Su and Lu ^[67]	✗	✓	✓	-
Gordo ^[21]	✗	✗	✗	-
Jaderberg et al. ^[30]	✓	✓	✗	490M
Jaderberg et al. ^[65]	✓	✓	✓	304M
Shi, Bai, and Yao ^[58]	✓	✓	✓	8.3M
Ours	✓	✓	✓	8.4M

5.6 中文场景讨论

现今的自然场景文本识别研究基本上都是基于英文和拉丁字母展开，对于中文场景下的文本识别研究内容甚少，且没有十分规范基准数据集对中文识别任务进行评估。为了检验本文算法在中文场景下的识别性能，我们使用了ICPR数据集^[68]进行实验和评估。该数据集为ICPR MTWI 2018挑战赛发布的公开数据集，数据来源主要是网络合成图像，以广告商标、产品描述为主。数据集的排版复杂，涵盖数十种字体，且有较多干扰背景，对文本识别算法来说是一个巨大的挑战。数据集经裁剪后包括训练集126384张，测试集15241张，裁剪后效果如图5-6所示。



图 5-6 数据集裁剪效果展示

5.6.1 实验设置及结果

实验中只使用了ICPR数据集的训练集，无额外训练数据的加入。经统计，总共需要预测的类别数为5709类，包括中英文字符、阿拉伯数字和部分特殊符号。实验参数设置与5.4.3小节中标准数据集实验参数设置相同，采用CRNN和HARN-base模型进行对比实验，收敛时间大概为30个小时。最终识别精度达到60.10%，高出CRNN模型58.14%将近2个百分点，但是相比于ICPR MTWI 2018挑战赛中获得第一名的方法仍然有一定差距。

5.6.2 实验结果分析

HARN模型在中文场景下的识别结果如图5-7所示。



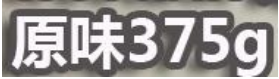





 prediction: “双滚珠轴承”  prediction: “净含量: 10ml”  prediction: “原味 375g”  prediction: “临时停车卡”	 prediction: “dbolo?”  prediction: “山竹千”  prediction: “丽丽老特”  prediction: “注音全彩手绘版”
识别正确的例子	识别错误的例子

图 5-7 中文数据识别效果

从实验可以看出对于大部分简单的数据样本能够准确识别，但是对于复杂多变的汉字结构模型的辨别能力仍然不够，形近字之间易混淆，例如将“山竹干”识别成了“山竹千”。对于一些特殊字符，如注册符号“®”难以识别。模型还存在漏识别现象，如“注音·全彩·手绘版”文本中汉字全部识别正确，但是符号“·”漏识别。分析其原因，类似于该类特殊字符的样本在训练集中的样本数太少，导致样本不均衡，模型未能充分学习到字符特征，识别性能不佳。此外，模型对于长文本的识别能力也稍显不足。^[68]中对数据集进行了字符串长度的统计，如图5-8所示。数据集中大部分文本长度集中在1-10个字符之间，但是超过10个字符的文本

仍然占据一定比重。对于长文本的识别也是自然场景下的文字识别算法需要攻克的难点和努力的方向。

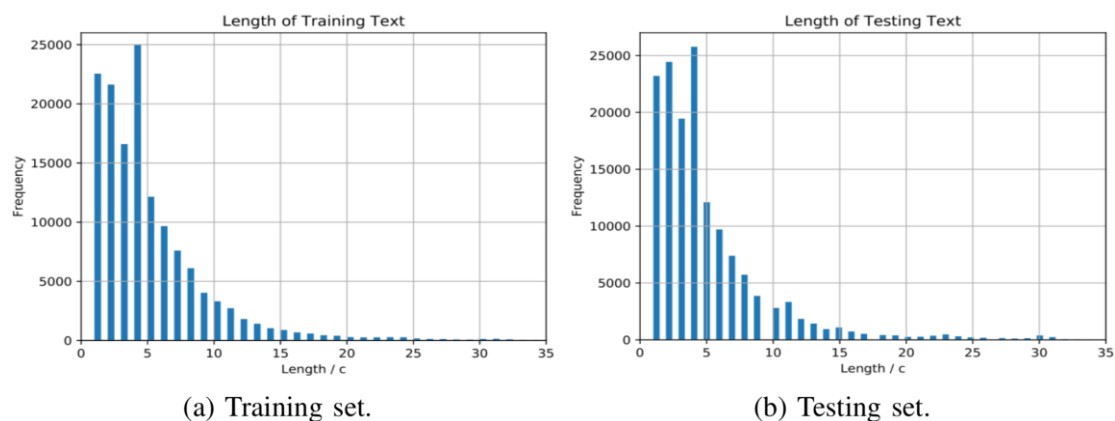


图 5-8 字符串长度统计

5.7 本章小结

本章主要介绍 HARN 方法的实验部分，首先介绍了实验中选取的训练集和测试集。然后介绍了实验之前的图像预处理，使图像更适用于网络的输入。接着介绍了实验的设置，包括实验的环境搭建，实验中使用的训练策略等。然后展示了几组不同实验的结果，并从可视化、泛化能力和综合能力等角度解析了实验结果，证明了本文提出的 HARN 方法具有较强的泛化性和鲁棒性。最后讨论了本文算法在中文场景下的识别性能。

第六章 总结与展望

6.1 本文总结

自然场景文字识别自提出以来一直备受关注,一方面其在工业生产中具有极强的实际应用价值,另一方面在学术领域也有着极高的研究意义。本文在前人研究的基础上,对现有的自然场景文字识别技术进行了深入剖析,根据现有技术的缺点与不足,提出了一套基于深度学习和神经网络的自然场景文字识别方法。该方法采用编解码框架,并引入层次化注意力机制,将文字识别问题转化为序列生成问题,取得了较好的识别效果。

本文提出的层次化注意力机制自然场景文字识别算法HARN,通过提供Image-Level Attention的CAN网络以及BiLSTM的双重编码,能够更好地捕捉图像空间局部特征和序列上下文特征,使提取到的特征具有更强的特征表达能力。在图像解码阶段同样引入注意力机制,基于ASGN的文字解码器提供了Context-Level Attention,能够实现字符更为精准的定位与识别。同时,该算法方法相比于传统识别方法具有一系列优势,其模型训练简单,成本低,识别精度高,同时模型的参数量少,内存开销少。该识别方法还在IIIT5K、SVT、ICDAR2013和ICDAR2015标准数据集上取得了优秀的成绩,这也证明了该套方法在自然场景文字识别领域中的优越性。

6.2 工作展望

尽管本文提出的基于注意力机制的文字识别方法对于大部分自然场景的文字识别行之有效,但是面对复杂的自然场景,仍然有许多亟待解决的问题和需要挖掘的方向。因此未来的工作内容主要围绕以下几点展开:

(1) 现实场景中的环境难以预测且十分复杂,尤其存在大量不规则的文本图片,这是自然场景文字识别算法的又一难点,对现有算法和模型提出了新的挑战。算法需要更强的通用性能和泛化能力,以应对复杂多变的自然场景。

(2) 自然场景下的中文字符识别也是一个重要的分支任务。中文字符的识别和英文字符的识别存在明显不同,相比于英文字符,中文字符的结构复杂,种类繁多,对算法的特征提取能力以及鲁棒性提出了更高的要求。

(3) 目前提出的算法有着较高的时空复杂度,虽然模型已经简化,但是仍然需要消耗很多计算资源,这就带来了一定的局限性。未来的研究方向希望朝着轻量级网络的方向发展,更适用于移动互联设备。

参考文献

- [1] T. F. Cootes, C. J. Taylor. Statistical models of appearance for computer vision: World Wide Web Publication, 2004.
- [2] D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco, CA: W.H. Freeman and Company, 1982.
- [3] Fujisawa H. Forty years of research in character and document recognition—an industrial perspective[J]. Pattern Recognition, 2008, 41(8):2435-2446.
- [4] S. Mori, C.Y. Suen, K. Yamamoto, Historical review of OCR research and development, Proc. IEEE 80 (7) (1992) 1029--1058.
- [5] G. Nagy, At the frontiers of OCR, Proc. IEEE 80 (7) (1992)1093--1100.
- [6] Yamamoto, K., Yamada, H., Saito, T., & Sakaga, I. (1986). RECOGNITION OF HANDPRINTED CHARACTERS IN THE FIRST LEVEL OF JIS CHINESE CHARACTERS. In Proceedings - International Conference on Pattern Recognition (pp. 570–572). IEEE.
- [7] J. Ohya, A. Shio and S. Akamatsu. Recognizing Characters in Scene Images. IEEE Trans. on PAMI. 1994, 16:214-220
- [8] C. Lee and A. Kankanhalli. Automatic Extraction of Characters in Complex Scene Images. International Journal of Pattern Recognition and Artificial Intelligence. 1995, 9:67-82.
- [9] Y. Zhong, K. Karu and A. K. Jain. Locating Text in Complex Color Images. Pattern Recognition. 1995:1523-1236.
- [10] Zhou and D. Lopresti. Extracting Text from WWW Images. In Proceedings of ICDAR. 1997:248-252.
- [11] Wang K, Belongie S. Word spotting in the wild. In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Heidelberg, Germany: Springer, 2010. 591-604.
- [12] Wang K, Babenko B, Belongie S. End-to-end scene text recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 1457-1464.
- [13] R. Lienhart, W. Dffelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. Technical Report TR-98-009, Praktische Informatik IV,

- University of Mannheim, 1998.
- [14] T. Sato, T. Kanade, E.K. Hughes, et al. Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions. *Multimedia Systems*. 1999,7(5): 385-395.
 - [15] Liu C M, Wang C H, Dai R W. Text detection in images based on unsupervised classification of edge-based features. In: *Proceedings of the 8th International Conference on Document Analysis and Recognition*. Seoul, South Korea: IEEE, 2005. 610-614.
 - [16] Yao C Bai X, Shi B, et al. Strokekts: A learned multi-scale representation for scene text recognition[C]//*Computer Vision and Pattern Recognition (CVPR)*, 2014 IEEE Conference on. IEEE, 2014: 4042-4049.
 - [17] Alsharif O, Pineau J. End-to-end text recognition with hybrid HMM maxout models[J]. 2013.
 - [18] Goodfellow I J, Warde-Farley D, Mirza M, et al. Maxout Networks[J]. *ICML* (3), 2013, 28: 1319-1327.
 - [19] Bissacco A, Cummins M, Netzer Y, et al. PhotoOCR: Reading Text in Uncontrolled Conditions[C]// *2013 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2013.
 - [20] Jaderberg M, Vedakh A, Zisserman A. Deep features for text spotting[M]//*Computer vision-ECCV 2014*. Springer International Publishing, 2014:512-528.
 - [21] A. Gordo. Supervised mid-level features for word image representation. In *CVPR*, 2015. 2, 6, 7.
 - [22] Mishra A, Alahari K, Jawahar C V. Scene text recognition using higher order language priors[C]//*BMVC 2012-23rd British Machine Vision Conference*. BMVA, 2012.
 - [23] Novikova T, Barinova O, Kohli P, et al. Large-lexicon attribute-consistent text recognition in natural images[M]//*Computer Vision-ECCV 2012*. Springer Berlin Heidelberg, 2012: 752-765.
 - [24] Goel V, Mishra A, Alahari K, et al. Whole is greater than sum of parts: Recognizing scene text words[C]//*Document Analysis and Recognition (ICDAR)*, 2013 12th International Conference on. IEEE, 2013: 398-402.
 - [25] Rodriguez Serrano J A, Perronnin F, Meylan R. Label embedding for text recognition[C]// *Proc. BMVC*. 2013.

- [26] Perronnin F, Liu Y, Sanchez J, et al. Large-scale image retrieval with compressed fisher vectors[C]//Computer Vision and Pattern Recognition(CVPR), 2010 IEEE Conference on. IEEE, 2010: 3384-3391.
- [27] Almazdn J, Gordo A, Fornes A, et al. Word spotting and recognition with embedded attributes[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2014, 36(12):2552-2566.
- [28] Goodfellow I J, Bulatov Y, Ibarz J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks[J]. 2013.
- [29] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition[J].2014.
- [30] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2014: 1-20.
- [31] Baoguang Shi, Xinggang Wang, Pengyuan Lv, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. arXiv preprint arXiv:1603.03915, 2016.
- [32] Baoguang S, Mingkun Y, Xinggang W, et al. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.
- [33] Zhazhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, Shuigeng Zhou. AON: Towards Arbitrarily-Oriented Text Recognition. arXiv preprint arXiv:1711.04226,2017.
- [34] Liu W, Chen C, Wong K, et al. STAR-Net: a SpaTial attention residue network for scene text recognition[J]. 2016.
- [35] Liu W, Chen C, Wong K. Char-Net: A Character-Aware Neural Network for Distorted Scene Text Recognition[J]. 2018.
- [36] Cheng Z, Bai F, Xu Y, et al. Focusing Attention: Towards Accurate Text Recognition in Natural Images[J]. 2017.
- [37] Simon M. Lucas, et al. ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal on Document Analysis and Recognition. 2005,7:105-122.
- [38] Shahab A, Shafait F, Dengel A. ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images[C]// Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

- [39] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 Robust Reading Competition[C]// 2013 12th International Conference on Document Analysis and Recognition. IEEE Computer Society, 2013.
- [40] Yao C, Bai X, Liu W, et al. Detecting Texts of Arbitrary Orientations in Natural Images[C]// Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
- [41] Veit A, Matera T, Neumann L, et al. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images[J]. 2016.
- [42] https://blog.csdn.net/qq_36047533/article/details/88671790
- [43] Mnih V, Heess N, Graves A, et al. Recurrent Models of Visual Attention[J]. Advances in neural information processing systems, 2014.
- [44] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. 2014.
- [45] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. 2017.
- [46] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015.
- [47] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[J]. Computer Science, 2015.
- [48] Hermann K M, Kočiský, Tomáš, Grefenstette E, et al. Teaching Machines to Read and Comprehend[J]. 2015.
- [49] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks[J]. Studies in Computational Intelligence, 2008, 385.
- [50] Raffel C, Ellis D P W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems[J]. 2015.
- [51] Daniluk, Micha, Rocktäschel, Tim, Welbl J, et al. Frustratingly Short Attention Spans in Neural Language Modeling[J]. 2017.
- [52] Wojna Z, Gorban A, Lee D S, et al. Attention-based Extraction of Structured Information from Street View Imagery[J]. 2017.
- [53] Lee C Y, Osindero S. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2231-2239.
- [54] T. Judd, K. Ehinger, F Durand. Learning to predict where humans look[C]. In ICCV, 2009.
- [55] Wu Y C, Yin F, Zhang X Y, et al. SCAN: Sliding Convolutional Attention

- Network for Scene Text Recognition[J]. 2018.
- [56] Jie H, Li S, Albanie S, et al. Squeeze-and-Excitation Networks[J]. 2017, PP (99):1-1.
- [57] Lin Z, Feng M, Santos C N D, et al. A Structured Self-attentive Sentence Embedding[J]. 2017.
- [58] 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 & Machine Intelligence, 2016, 39(11):2298-2304.
- [59] A. Graves, S. Fernandez, F. J. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, 2006.
- [60] Tao W, Wu D J, Coates A, et al. End-to-End Text Recognition with Convolutional Neural Networks[C]// International Conference on Pattern Recognition. 2012.
- [61] <https://github.com/aleju/imgaug>
- [62] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015.
- [63] C.-Y. Lee and S. Osindero. Recursive recurrent nets with attention modeling for OCR in the wild. In Proceedings of Computer Vision and Pattern Recognition (CVPR), pages 2231–2239, 2016.
- [64] F. Yin, Y. Wu, X. Zhang, and C. Liu. Scene text recognition with sliding convolutional character models. CoRR, abs/1709.01727 (2017), 2017.
- [65] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. In ICLR, 2015.
- [66] C. Yao, X. Bai, B. Shi, and W. Liu. Strokelets: A learned multi-scale representation for scene text recognition. In CVPR, 2014.
- [67] B. Su and S. Lu. Accurate scene text recognition based on recurrent neural network. In ACCV, 2014.
- [68] M. He, Y. Liu, Z. Yang, S. Zhang, C. Luo, F. Gao, Q. Zheng, Y. S. Wang, X. Zhang, and L. Jin. Icpr2018 contest on robust reading for multi-type web images. 2018 24th International Conference on Pattern Recognition (ICPR), pages 7–12, 2018.

致谢

两年半的校园时光，转瞬即逝，计算机楼下的婆娑竹影，三星河畔的丝丝荷香，木拱桥上的点点星光，都在耳畔、鼻前、心尖。这几年在学习与生活上都遇到了前所未有的挑战，从看文献心浮气躁到发现突破点；从代码基础薄弱到独立设计算法。所幸在这段不那么轻松的时光里，有父母、老师的支持与开导，有校内外朋友的帮助与鼓励，我才有勇气一直坚持努力，顺利完成学业。

首先，我要感谢我的导师薛向阳教授。我很荣幸在研究生生涯能够遇见薛老师。他为人正直，教学严谨，对我们每个学生都有很高的期望和要求。同时，他在学术道路的选择上也能充分尊重学生们的想法。也正是薛老师让我能在学术道路上按照自己的想法发展，我才能以一个通信方向的背景成功转到计算机专业，并对计算机视觉领域做深入的探索。同时，我也要感谢我的另一个导师李斌教授。李老师知识渊博，平易近人，对待科研的态度让我深深折服。从他身上，我感受到了对做研究的热情与严谨，对做科研的执着与追求。感谢两位导师对我科研上的帮助与关心，是你们的无私奉献为实验室提供了良好的科研环境和氛围。我还想感谢冯瑞老师，是您把我领入复旦大学的校门，让我有幸成为复旦莘莘学子中的一员。另外还要感谢我在高研院的两位导师，叶浩博士与郑莹斌博士，在高研院的学习和经历，为我的科研道路打下了坚实的基础。

感谢我尊敬的师兄马建奇，每每在我遇到难题之时，总会耐心讲解，将自己对于科研方向的探索与擅长的各种技能倾囊相授。也感谢我的师姐王丽、王晓梅，从与你们的交谈中，我感受到了实验室互帮互助，轻松愉悦的氛围。我还想感谢我可爱的师弟师妹们，我会永远记得打比赛时与大家并肩作战的日子。同时，我也感谢同门们的鼎力支持，通过与你们在科研上的交流，我得到了长足的进步。

最后，我尤其要感谢我的父母。从咿呀学语的蹒跚小儿到落落大方的花季少女，多年以来，他们在物质上无怨无悔地支持，生活上无微不至地关怀。儿行千里母担忧，我自远方求学，亦难想象家中父母对我的牵挂。父母日渐老去的身影，让我心怀愧疚。以后的人生道路上，让我来做你们的大树，为你们遮风挡雨。

读书十九载，得之于人者甚多而出之于己者甚少。那个夏天，收到复旦大学的录取通知书，乃是我一生之骄傲。做一个自由而无用的灵魂，我不能保证自己未来人生的轨迹永远不会出现偏移，但博学而笃志，切问而近思这八个字，我将永远铭刻在心中，念兹在兹，薪尽火传。感谢在学校每一位帮助我的老师、同学和朋友们，还有我一对敬爱的父母，文及此处，行将收尾，字数有限而不能尽表感谢之意。

复旦大学 学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：_____ 日期：_____

复旦大学 学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____