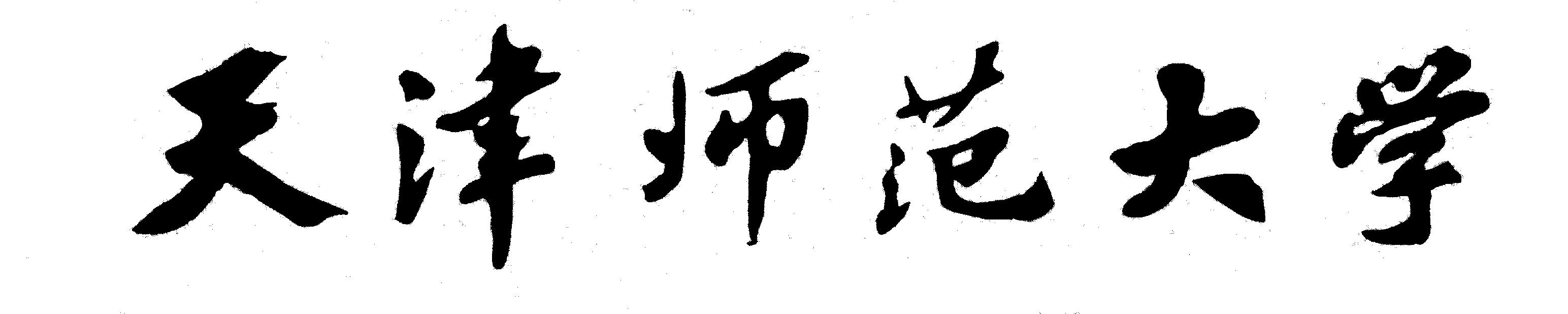
分类编号：单位代码：10065

密 级： 学 号：1110090016



研究生学位论文

论文题目： 基于RNN的端对端场景文本识别方法的研究

学生姓名： 安兴乐 申请学位级别： 硕士

申请专业名称： 计算机应用技术

研究方向： 图像处理与模式识别

指导教师姓名： 朱远平 技术职称： 副教授

提交论文日期： 2016 年 12 月 10 日

原创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研

究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其

他人已经发表或撰写过的研究成果，也不包含为获得天津师范大学或其它教育

机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何

贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名： 日期： 年 月 日

研究生学位论文使用授权说明

本人完全了解天津师范大学关于收集、保存、使用研究生学位论文的规定，

即：

* 按照学校要求向图书馆提交学位论文的印刷本和电子版本；
* 图书馆有权保存学位论文的印刷本和电子版，并通过校园网向本校读者提供全文与阅览服务。
* 图书馆可以采用数字化或其它手段保存论文；
* 因某种特殊原因需要延迟发布学位论文，按学位论文保密规定处理，保密论文在解密后遵守此规定。

论文作者签名： 导师签名：

日期： 年 月 日

## 天津师范大学硕士学位论文原创声明

本人郑重声明:此处所提交的硕士学位论文《基于RNN的端对端场景文本识别方法的研究》，是本人在导师指导下，在天津师范大学攻读硕士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名： 日期： 年 月 日

## 天津师范大学硕士学位论文使用授权书

《基于RNN的端对端场景文本识别方法的研究》系本人在天津师范大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归天津师范大学所有，本论文的研究内容不得以其他单位的名义发表。本人完全了解天津师范大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅，同意学校将论文加入《中国优秀博硕士学位论文全文数据库》和编入《中国知识资源总库》。本人授权天津师范大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

本学位论文属于（请在以下相应方框内打“√");

保密□,在 年解密后适用本授权书

不保密□

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 作者签名 | 日期 | 年 | 月 | 日 |
| 导师签名: | 日期: | 年 | 月 | 日 |

: :

# 摘 要

随着移动智能设备的普及，图像的获取越来越便捷，提取并理解智能设备拍摄的图像中的信息具有重要的应用价值。文字由于包含了大量丰富的语义信息，对于计算机理解图像具有重要意义。文本识别作为计算机视觉领域基础性研究工作，自然场景的文本识别具有深远的研究意义和实用价值。在图像检索、人机交互、车牌识别等领域有着广泛的应用。自然场景由于复杂的背景、多样化的字体、以及随意的分布等问题，使得传统的光学字符识别技术（OCR）难以被广泛应用。自然场景中的文字检测与文本识别仍然存在很多需要解决的技术难点。

得益于计算机硬件性能的提高，以及大数据获取越来越便捷，深度学习技术近年来取得了重大突破，在模式识别领域被广泛应用。本文借鉴图像检索、目标检测、自然语言处理等领域的进展，结合深度学习技术，提出了新的识别方法，本文主要工作如下：

1. 提出了基于CNN和RNN的文本识别方法。利用CNN提取原始图像的抽象特征，然后将特征送入RNN中建模图像全局信息。针对传统的文本识别需要较深的专业背景知识，文本检测与识别等过于复杂等问题，该方法采用端到端训练，无需过多的背景知识。将检测与识别融为一体。大大降低了训练难度。
2. 对自己生成的验证码图片数据集和Google的街景门牌号码SVHN数据集上进行了两部分实验。第一部分实验验证了单纯RNN对文本识别的有效性；第二部分实验探索CNN结合RNN对文本识别的提高以及不同的滑动窗对识别率的影响。

关键字：文本识别，深度学习，RNN，CNN

目录

1. 绪论
   1. 自然场景文本识别的研究背景及意义
   2. 自然场景文本识别难点分析
   3. 本文的主要内容与结构安排
2. 循环神经网络
3. 基于RNN的文本识别方法
4. 实验系统设计与总结
5. 绪论
   1. 文本识别的研究背景及意义

文字是人们重要的交流工具，如何让计算机来认识并处理人类的文字一直是计算机视觉与模式识别领域重要的课题。早在上个世纪初期光学字符识别（Optical Character Recognition, OCR）技术就被提出，经过长期发展，OCR在工业领域的应用日臻成熟。但是OCR技术要求文档图像具有清晰简单的背景，较高的分辨率，而自然场景中的图像往往没有鲜明的对比背景，光照变换非常大，分辨率较低等使得OCR技术在自然场景中几乎无法使用。

然而随着智能移动设备走入千家万户，在智能移动上直接理解并处理文字信息必然有着巨大的应用需求。相对于多媒体文档中的常见的亮度、色彩、形状等底层信息，文字能够直接传递高级语义信息。合理地利用多媒体文档上的文字信息，能够极大的推动以下领域的发展：

1. 图像与视频检索。移动互联网时代图像与视频文件越来越多，如何对海量的图像与视频进行检索是一个重要问题。基于内容的图像和视频检索的研究目的是根据图像和视频的抽象信息，根据一定的相似性指标来进行检索。目前难点在于计算机对图像与视频的相似性判断是根据图像的底层视觉特征来进行的相似性判断。而人由于在生活中拥有大量的经验与知识，对图像的相似性判断是根据语义信息来进行的相似性判断，这就产生了人对图像所理解的“语义相似”与计算机理解的“视觉特征相似”的“语义鸿沟”问题。而根据图像或视频中所识别的文本信息，可以提供较为丰富的高层语义信息，进而可以根据这些语义信息来对图像或者视频进行检索，或者依据语义信息对图像或者视频进行分类。
2. 车牌识别。随着人们生活水平的提高，机动车数量也越来越多。高效的车辆管理不仅可以缓解交通压力，而且对公共安全有重要意义。车牌往往就是唯一确定车辆身份的标志。现在往往在重要路口都有摄像头监控过往车辆的车牌，记录车辆的行驶信息，可以对违反交通法规的车辆记录在案，或者对可疑车辆及时报警。总之，车牌识别系统对车辆管理具有重要意义，是模式识别与计算机视觉领域重要课题之一。
3. 无人驾驶。随着传感器与自动导航技术的发展，自动驾驶汽车的也开始走入人们的视线。2012年5月Google的第一辆自动驾驶汽车获得自动驾驶许可证。自动驾驶可以有效的辅助甚至代替人类操作机动车量，能够为人类节省不少的驾驶时间，并且可以极大的减少能源消耗，降低交通事故的发生，让乘坐车辆更加安全。高速行驶车辆在复杂路况下识别道路指示牌和建筑标志中的文字显得更加重要。
   1. 自然场景文本识别难点分析

近几十年来文字识别研究得到了空前的发展，一大批科研工作者致力于这方面的研究，但是目前仍然不存在一种在所有自然场景中适用的实际方法来识别文本。如图1-1所示，光照不均匀，分辨率过低，视角 

图1-1 自然场景下的文本图像示例

的多变性，拍摄设备的移动都会导致图像模糊，文字难以识别。传统应用于扫描文档的OCR引擎几乎无法正确的扫描识别出这样的文字。现在越来越多的研究将机器学习方面的算法引入到场景文本识别领域来，取得了很好的效果，但是距离实用化还有一段距离。为取得更好的识别效果，训练算法需要大量的专业领域知识来结合机器学习，对科研工作者提出较高的要求。

综上所述，随着智能移动终端设备以及移动互联网的逐渐普及，多媒体文档中文字识别会越来越重要，而目前的自然场景文字识别技术依然不能满足现实需要，所以展开对文字识别的研究显得非常迫切。

1.3 本文的主要内容与结构安排

本文针对上述的问题展开探索与研究，提出了基于深度学习的文本识别方法，以实现端到端的训练，无需专业领域的背景知识。

本文的主要研究任务如下：

1. 提出了结合CNN与RNN的新型网络结构，实现了端到端的文字识别。
2. 构建一个标准验证码数据集，对本文模型进行评估。

本文的内容与结构安排如下：

第二章：循环神经网络RNN。对人工神经网络（Artificial Neural Networks），递归神经网络(Recurrent Neural Network)的基本理论进行简要的介绍。

第三章：基于RNN的文本识别方法。结合CNN与RNN的新型网络结构来识别文本。

第四章： 全文总结与展望。本文算法的提高与不足进行总结，提出未来要开展的工作作出展望。

## 循环神经网络

### 传统人工神经网络

人工神经网络（Artificial Neural Networks, ANN）简称神经网络，是模拟动物的神经网络的计算模型。这些计算模型通过模拟动物大脑的神经网络进行抽象计算，构建人工神经元，然后按照一定的拓扑结构来构建神经元之间的连接，来模拟动物的神经网络。1949年，生理学家D.O.Hebb提出改变神经元连接强度的Hebb规则，为神经网络的学习算法奠定了基础。这时的神经网络实只有一个隐含层，且受限于当时计算机硬件资源，并没有预期的效果。1974年，哈佛大学的博士生Paul Werbos提出了著名的BP算法(back—

propagating，反向传播法)，但没有引起学界的关注。但在1986年，BP算法

受到多伦多大学的Geoffrey Hinton等人的关注，他们并在((Nature))上发表了一

篇对这个算法的研究，这使得人工神经网络再次得到关注。

#### 2.1.1 神经元

神经元是神经网络最简单的组成部分。人工神经元和感知器类似，模拟动物的神经元特性，接受一组输入并产生输出。下图是一个最简单的神经网络模型，它只包含一个神经元，如图所示：

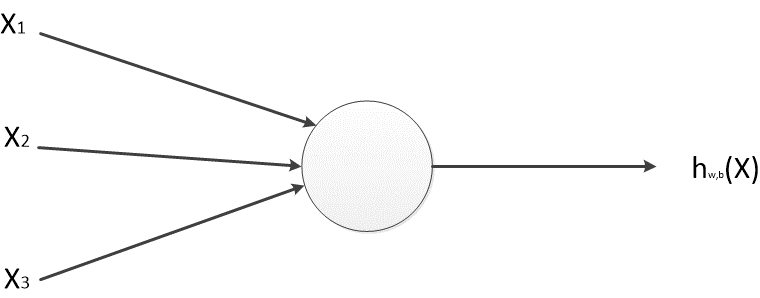


图1-2 神经元结构

这个神经元的输入为X1， X2， X3，输出为hw,b(X)= f（WTx）,f为激活函数，W为线性变换，b为偏置项。

#### 2.1.2 常用激活函数

线性模型的表达能力往往有限，为了增强神经网络的表达能力，我们需要使用连续非线性激活函数。下面介绍几种神经网络常见的激活函数。

Logistic函数

Logistic函数是一种sigmoid型函数，定义为：

Logistic函数将输入映射到（0，1）范围内。函数的特点和动物神经元类似，对一些输入会产生兴奋，对另外一些输入会产生抑制。当输入越大时，函数输出越接近1；输入越小时，函数输出越接近0.

Tanh函数

Tanh函数也是一种sigmoid型函数，定义为：

tanh函数可以看作是放大并平移后的logistic函数：

tanh(x) =

图1-3给出了logistic函数与tanh函数的形状。

线性修正单元

线性修正单元（rectified linear unit, ReLU）也称为rectifier函数。在深层神经网络中普遍被使用。ReLU形状如图1-4所示，是一个“斜坡”函数，定义为：

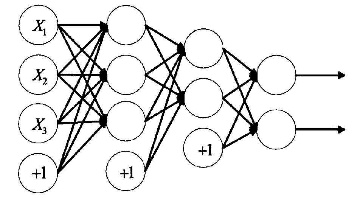
rectifier(x) =

采用ReLU函数的神经网络在计算上更加高效。Relu函数会使一部分神经元的输出为0，这样就造成了网络的稀疏性，并且减少了参数的相互依存关系，缓解了过拟合问题的发生。

2.1.3 前馈神经网络

单层神经元通过训练只能得到一个线性分类面，其分类能力有限。动物大脑中神经细胞是相互连接在一起的，我们可以构建类似的神经网络结构，组成多层的人工神经网络。多层神经网络能够表示多个非线性分类面，分类能力大大增强。其中，前馈神经网络是最简单的多层神经网络。

在前馈神经网络中，各神经元分属于不同的层。每一层的神经元接收上一层神经元输出的信号，并输出给下一层的神经元。整个网络中无反馈，信号从输入层向输出层单向流动，可以用一个有向无环图表示。如图1-3所示是一种广泛使用的前馈神经网络。它包含输入层 L1，两个隐藏层L2,L3,以及输出层L4。



Layer L1 Layer L2 Layer L3 Layer L4

图1-5 三层神经网络

我们使用下面的记号来描述前馈神经网络。

L: 表示神经网络的层数；:表示第层神经元的个数；:表示L层神经元的激活函数；:表示L-1层到第L层的权重矩阵；:表示L-1层到第L层的偏置；:表示L层神经元的输入；:表示L层神经元的输出。我们可以得到前馈神经网络第L层神经元的输入为：

第L层的神经元输出为：

公式合并为：

整个网络可以看作一个函数。

### 2.2 循环神经网络

#### 2.2.1 简单循环神经网络

传统神经网络中每层之间的神经元是没有连接的，这种网络对于很多问题无能为力。循环神经网络（Recurrent Neural Network, RNN ），利用带自反馈的神经元可以处理任意长度的输入序列。这种网络更加适合处理与序列有关的任务。它已被广泛应用于语音识别，自动翻译，聊天机器人等任务上。

一个简单循环神经网络按时序展开之后如图1-6 所示：

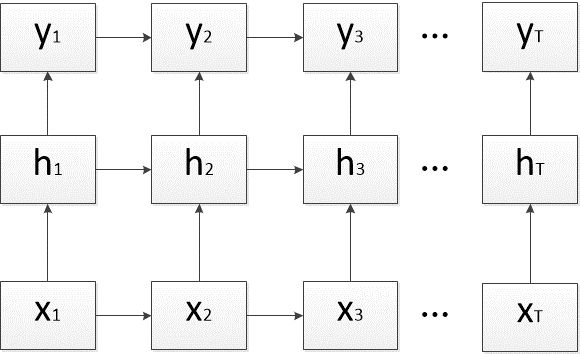


图1-6 循环神经网络按时序展开

这种网络在神经元之间既有内部的反馈连接，又有前馈连接。假设在时刻t时，输入为xt, 隐层状态为ht。我们可以得到如下的表达形式：

ht = f(Uht+Wxt+b) (2.1)

f为激活函数，U和W为线性变换函数，b为偏置项。我们可以看到ht不仅和当前时刻的输入有关，也和上一个时刻的隐层状态有关。

#### 2.2.3 RNN反向传播算法

#### 循环神经网络的 参数训练可以通过随时间进行的反向传播算法来确定。假设循环神经网络在每个时刻t都有一个损失。整个序列损失为 。序列的损失J关于U的梯度为：

⑵

⑶

根据链式法则，展开可得

, ⑷

将项展开， ⑸

. ⑹

令 ƞ 。

将ƞ代入公式⑷，可得

显然，当ƞ>1，时，，会出现梯度爆炸问题。而如果ƞ>1，时，，会出现梯度消失。因此，在实际中简单循环神经网络只能学习到短周期的时序依赖关系，即长期依赖问题。

#### 2.2.4 LSTM

为解决这个问题, Hochreiter 和 Schmidhuber 引入了门机制（Gating Mechanism）来控制信息的累计速度，并选择性遗忘以前的信息。这就是长短时记忆神经网络（Long Short-Term Memory Neural Network, LSTM）。在时刻t时候，记忆单元Ct记录了到时刻

t 为止的所有历史信息，受三个门单元控制：输入门it，遗忘门ft和输出门Ot。LSTM的更新方式为：

(2.2)

(2.3)

(2.4)

(2.5)

(2.6)

(2.7)

这里是logistic函数，是对角矩阵，是当前时刻的输入。遗忘门控制每一个内存单元需要遗忘的信息，输入门控制每一个内存单元记忆多少新的信息，输出门控制每一个内存单元输出多少信息。



图2-7 LSTM 结构图

LSTM网络可以学到长周期的历史信息。LSTM网络的计算机构可以用 图2-7来表示。

## 第三章 基于RNN的文本识别方法

3.1 问题分析

计算机对文本的识别是通过将图像信息转化为计算机可表示的信息并进行处理的过程。由于自然场景的文字及其复杂，现在的识别方法没有很好的鲁棒性，并且需要较深的专业领域知识，通用性较差。 同样的问题，在物体分类、图像分割等其他计算机视觉领域也存在。随着深度学习理论的发展，卷积神经网络在计算机视觉领域取得了不错的成绩。卷积神经网络直接输入图像像素信息，不需要人工设计特征，同时也大大减少了数据预处理的工作量。

虽然基于卷积神经网络的自然场景文字识别取得了一定的成果，但是仍然存在一些不足。卷积神经网络针对单字的识别虽然较高，但是对于复杂的自然场景下的文本信息，却依然需要配合传统的字符定位与分割方法，不能实现端到端的训练。而且当图像所含文本较长时，识别率会明显下降。

我们借鉴深度学习在语音识别上取得的成就，将文本识别任务作为特殊的“语音信息”进行处理。本章将会结合CNN与RNN，提出一种新型网络结构，用于文本识别，以期能够实现端到端的自然场景文本识别。

3.2 图像预处理

自然场景图像虽然大部分是彩色图，但是实验发现图像转换为灰度图并不影响识别精度。为了减小计算量，原始图像可以按照灰度图来读取。为了减少光照的影响，需要对图像进行灰度值归一化。

为了后续神经网络处理方便，图像的输入需要保持固定大小。这里将图像设为约60 像素。图像预处理示意图如下所示：

C:\Users\a\AppData\Local\Microsoft\Windows\INetCacheContent.Word\13.png C:\Users\a\AppData\Local\Microsoft\Windows\INetCacheContent.Word\13_42.bmp

图3-1 图像预处理

3.3 方法概述

本文提出基于改进的RNN为基础框架。整个识别流程如图3-2所示：

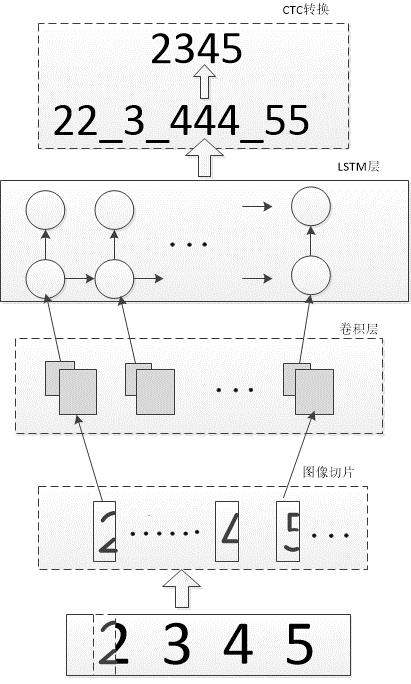


图3-2 文本识别流程图

训练集中图片归一化为6060像素，使用滑动窗经过图像，得到若干组图像切片。将切片送入卷积层提取特征，得到切片的若干组特征，将切片特征经过全连接网络后送入LSTM网络。网络输出经过CTC层转换后得到最终结果。

3.4 卷积层

卷积神经网络（Convolutional Neural Networks, CNN）属于前馈神经网络，通过局部感受野、权值共享和下采样实现了对输入图像的位移变化、尺度变化、形变变化的不变性。对于卷积神经网络第L层神经元的输入有如下定义：

这里为第L层滤波器，代表L-1层神经元的输入，代表偏置矩阵。f为激活函数。通常会使用多组滤波器来得到多组输出，以此来增强卷积层的表达能力。滤波器可以看作特征提取器，每一组输出可以看作图像经过特征提取后得到的特征。

通常在卷积后加上一个池化（Pooling）操作，也称为下采样（Subsampling）。下采样可以显著降低特征维数，避免过拟合。

相对于RNN网络强大的时间序列建模能力，CNN网络更擅长对图像的空间关系进行建模。卷积层的作用是提取图像局部区域的特征，以便LSTM网络进行预测。

3.5 Connectionist Temporal Classification

在LSTM网络中，如果序列的长度为T，则会产生T个输出。这就需要将LSTM的输出解码为正确的类别。Connectionist Temporal Classification (CTC)技术用来实现LSTM网络输出与数据标签对齐，计算LSTM网络训练过程中标签与LSTM网络输出之间的Loss。图3-3 是CTC对LSTM网络输出到数据标签的路径搜索示意图。

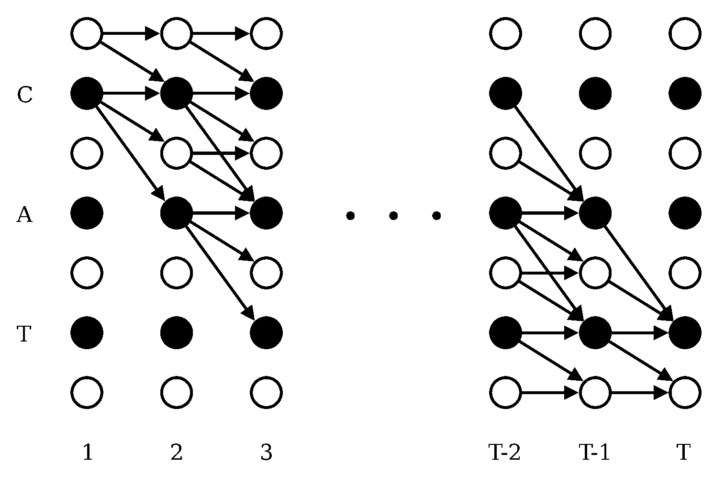


图3-3 CTC路径搜索示意图

通过图3-3可以发现，CTC在数据标签中增加空白标签来分割标签之间的间隔，最后会把空白符号和预测出的重复符号消除。假设输入为。CTC技术使得模型从输入端输入数据后可以在输出端直接取得识别结果，实现了端到端训练算法。

第四章 实验系统设计与总结

深度学习的训练要求计算机有较高的计算能力。2012年ImageNet竞赛第一名AlexNet模型，使用两块GTX 580 GPU,训练时间长达六天，2014年ImageNet竞赛第一名 Oxford VGGNet模型，使用4块Titan Black GPU，训练时间更是长达3周。另外，神经网络所含的参数众多，如神经网络层数、神经元数量、规则化系数、激活函数等。训练时超参数如学习率，批尺寸（batch size）大小等也会明显影响收敛速度。

本章尝试使用一些策略来缓解这些问题，并探索不同的参数对实验的影响。

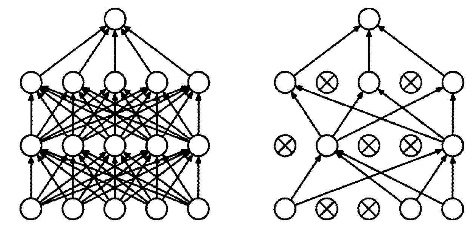
4.1 基于GPU的训练

GPU相对CPU有更多的计算核心，每个核拥有相对较小的缓存，少而简单的数字逻辑运算但与。这使得GPU更加适合处理相对简单而重复的计算任务。目前主流的深度学习研究工作越来越多的使用GPU来对训练过程进行加速。训练过程中采用 批随机梯度下降中，每批样本的数量更具GPU显存大小进行合理设置，尽可能充分利用GPU显存。

本实验使用Nvidia GTX 980ti GPU，深度学习框架Tensorflow基于cudnn加速。为减少CPU与GPU通讯耗时，一次性将全部训练数据读入内存，每一轮次迭代由内存送入GPU数据。这样大大增加了GPU利用率，减少资源浪费，加速训练过程。

4.2 Dropout

Dropout 作为一种防止过拟合的方法，广泛被用在神经网络训练中。训练时，随机对全连接层直接连接移除。



a. 不进行dropout的全连接网络 b. dropout的全连接网络

图 4-1 dropout示例

Wojciech Zaremba在2015年将dropout应用与RNN网络，取得了不错的效果。图4-2是RNN中dropout应用示例：

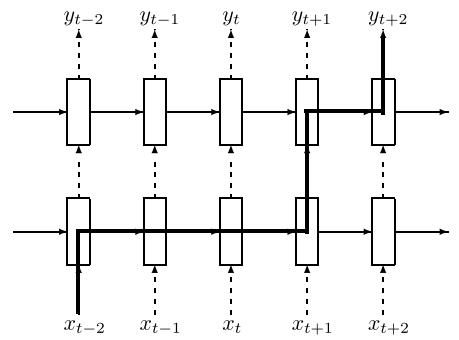


图4-2 dropout在RNN中应用

图中虚线代表应用dropout的连接。单步的层间连接进行一定概率的dropout，序列的循环连接不使用dropout。

4.3 实验系统设计

4.3.1 数据集

本文使用了两种数据集，根据python验证码库*captcha生成的验证码图片以及被广泛使用的Google 街景门牌号数据集（SVHN）。*

C:\Users\a\AppData\Local\Microsoft\Windows\INetCacheContent.Word\16_6fd9.bmpC:\Users\a\AppData\Local\Microsoft\Windows\INetCacheContent.Word\28.png

1. *Captcha生成的验证码图片 b. SVHN数据*

*图4-3 数据集*

*Captcha生成的验证码噪声、背景图与字符数均可以由用户决定，我们这里生成长度为4个与3个字符的验证码图片。SVHN数据集是*真实世界的街道门牌号数字识别数据集。SVHN有10类数字，长度不等。其中73257幅图片用于训练，26032幅图片用于测试，531131幅图片是较容易的图片用于额外的补充数据。

4.3.2 实验环境

实验使用python进行数据预处理，使用tensorflow框架进行神经网络训练。具体实验环境如表 4.1 所示。

表4.1 实验环境

|  |  |
| --- | --- |
| *CPU* | *Intel core i7-4790K* |
| *GPU* | *NVIDIA GTX980ti* |
| *内存* | *16GB* |
| *操作系统* | *Ubuntu 14.04* |
| *开发语言* | *Python* |
| *GPU开发库* | *CUDA 5.1* |
| *图像处理库* | *Opencv 3.1* |
| *深度学习框架* | *Tensorflow 0.9* |