

# 机器学习纳米学位

## 开题报告

Xuefeng Sun 优达学城  
2018年7月21号

### 项目背景

光学图像识别 OCR (Optical Character Recognition) 领域，是通过扫描等光学输入方式将各种票据、报刊、书籍、文稿及其它印刷品的文字转化为图像信息，再利用文字识别技术将图像信息转化为可以使用的计算机输入技术。可应用于银行票据、大量文字资料、档案卷宗、文案的录入和处理领域。OCR技术已经有了很长的历史，并且比较成熟。被广泛应用到车牌号，门牌号，自动身份证识别。利用机器自动识别，可以快速准确的将图片中包含的信息转化为文字输出。代替人工识别，节省了大量的人力。深度学习是近年来非常热门的一类机器学习，在图像识别方面拥有很高的识别度和准确度。本项目也因此结合深度学习技术对图片进行识别，正确提取图片中的算式表达式。图像序列识别旨在提取和分析图像中的序列，一般可以被分为长度固定的序列，例如：身份证号，也有长度不固定的序列，例如：文字提取。本项目是一个算式识别项目，其序列的长度也是不固定的。

### 问题描述

本项目将原始图片作为输入，输出为图片中的算式表达式。如图：




图片由数学表达式和噪点组成，每一个字符都有可能旋转或者和其相邻的字符粘连在一起。目前比较流行的识别的方法有分割法识别和深度学习识别。分割法，比较传统就是就是将每一个字符“扣”下来，再利用KNN/SVM 等这些机器学习算法进行单个字符的识别。例如：



深度学习识别是利用了卷积神经网络 CNN (Convolutional Neural Network) 提取图片特征和循环神经网络 RNN (Recurrent Neural Network) 处理图片中的序列自动的提取图片特征并且识别图片中的序列。通过这样的一个模型自动的识别和分析图片，让机器能够自我学习如何识别。

### 数据和输入

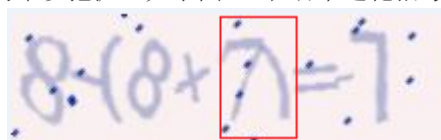
项目所用的数据集来源于：<https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd009/MLND+Capstone/Mathematical Expression Recognition train.zip> 整个数据集有10万张彩色图片和一张对应于每张图片的标签列表。每一张图片都包含一个算式。标签列表中与之文件名相对的则是该图片对应的算式表达式。例如：

标签列表		图片
filename	label	0.jpg
train/0.jpg	(0+0)+9=9	
train/1.jpg	9*8+6=78	

算式中的字符为数字0-9，括号()，以及运算符\*+- 和 =。每张图片中算式序列长短不一，最短的序列为7，最长是11。由于图片中字符的旋转，字符之间的粘连，以及图片中的干扰物都给图片识别带来了一定的难度。在该项目中将整个数据集（10万张图片）划分为训练集，用于训练模型参数；验证集，对模型进行验证；测试集，用于最终的模型测试所用数据。

## 解决方法描述

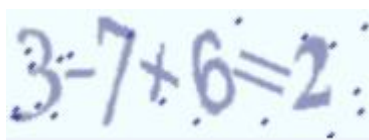
传统的解决思路比较暴力，先对图片进行去噪，二值化；然后切分字符；再对每一个字符单独识别。这种方法虽然简单常用，但是也会有很大的困难，主要表现在：1）去噪，数据集中每一个图片都有一些噪点，并且数字有旋转和变形。比如 \* 和 + 非常的相近，这也会给识别带来很大难度，阻碍识别的准确性。2）数字粘连，这是最难处理的一个问题，如何准确的切分，非常的难以把握。如下图：7和右半边花括号粘连。



基于以上传统方法的缺点，本项目将构建一个 CRNN (Convolutional Recurrent Neural Network) 模型。深度学习非常擅长端到端的学习，结合卷积神经网络 CNN 和 循环神经网络 RNN 构建的端到端的模型，有几个特点：1）不需要切分字符，能够直接通过标签对图像进行学习；2）有RNN 的特性能够学习识别一个序列；3）不需要太过于关系序列长度；4）有很高的准确度。

## 评估标准

本项目将选择准确度作为评判标准，只有当所有算式字符都正确，才判为该识别图片被正确识别，反之识别错误。例如：



，只有输出为“3-7+6=2”是才能记为正确。

## 基准测试

项目要求在测试数据集上的准确度为99%以上。

## 项目设计

整个项目由以下几部分组成：

### 1. 数据分析和处理

1. 将10万张图片按照（8:1:1）比例进行分割为训练集，验证集和测试集；
2. 每张原始图片为300 x 64 x 3，将其进行缩放为150 x 32 x 3，加快训练的过程；
3. 由于数据量大，一次性加载所消耗的内存比较大，所以构建一个生成器分批次加载数据。对于每一个图片而言是一个不定长的序列，现将其转化为定长的序列，对于长度不足最大长度的序列以某一数据进行填充。

2. 模型构建和模型训练

- 1. 构建卷积神经网络 CNN 模型对原始图片进行特征的提取；
- 2. 以 CNN 模型的输出作为循环神经网络 RNN 的输入构建一个端对端的模型；
- 3. RNN 输出输入到 CTC 算法计算 loss 并进行序列化预测；
- 4. 自定义评估函数；
- 5. 训练模型。

3. 可视化结果和测试

- 1. 以图表形式可视化模型 loss 和准确率，并根据结果微调参数重建模型；
- 2. 在测试集上进行验证，并获得99%的准确率。

4. 模型结构以及参数调优

模型结构如下。本结构是参照“参考文献2”中模型结构做了一些调整。

Type	Configuration
Input	image size: (150, 32, 3)
Convolution	kernels 64, size:(3, 3), s:1, p:"valid"
MaxPooling	size: (2, 2), s:2
BatchNormaliztion	
Convolution	kernels 128, size:(3, 3), s:1, p:"valid"
Convolution	kernels 128, size:(3, 3), s:1, p:"valid"
Convolution	kernels 128, size:(3, 3), s:1, p:"valid"
MaxPooling	size: (2, 2), s:2
BatchNormaliztion	
Convolution	kernels 256, size:(3, 3), s:1, p:"same"
Convolution	kernels 256, size:(3, 3), s:1, p:"same"
Convolution	kernels 256, size:(3, 3), s:1, p:"same"
Convolution	kernels 256, size:(3, 3), s:1, p:"same"
MaxPooling	size: (1, 2), s:1
BatchNormaliztion	
Convolution	kernels 512, size:(3, 3), s:1, p:"same"
Convolution	kernels 512, size:(2, 2), s:1, p:"same"
Convolution	kernels 512, size:(2, 2), s:1, p:"same"
MaxPooling	size: (1, 2), s:1
BatchNormaliztion	
Reshape	
Dense	128
Dropout	0.2
Bidirectional(GRU)	rnn_size:128
Bidirectional(GRU)	rnn_size:128
Dropout	0.3
Dense	n_class:(16 + 1)
ctc	
optimizer	adadelta

针对以下几种可能的情况对参数进行调优：

### 1. 欠拟合

- 增加卷积层
- 增加训练次数
- 降低 Dropout
- 数据增强

### 2. 过拟合

- 增加 Dropout
- 正则化

### 3. 模型不收敛或者收敛慢

- 更换 optimizer
- 调整学习速率
- 更换 activation function

## 参考文献

1. <https://baike.baidu.com/item/OCR%E6%8A%80%E6%9C%AF/15695472?fr=aladdin>
2. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition