亿保杯–票据分割比赛

<Excerpt in index | 首页摘要>
这只是浙大的一个校内比赛
KeyWords Plus:　关于票据重叠文本的情况进行分割

- **relevant**：出发点：很多票据会出现打印字错位，甚至是打印字和底板重叠，利用分割算法进行分割重叠文本，以便于后期的单据识别。
- **coding**：Github

<The rest of contents | 余下全文>

简介

初赛十分简单，就是用分割网络对重叠文本进行分割即可，评价指标是IOU

智慧E保 jianbaolife.com

亿保杯算法设计大赛

课题背景

上海亿保作为全国领先的 TPA 服务公司，致力于推进票据分类录入的全流程自动化，其中我们需要对发票中的文字进行检测和识别。有很多票据会出现打印字错位，甚至是打印字和底板重叠，一般的机器学习方法不能准确高效的处理这些票据。所以我们希望在不影响检测识别情况下能够将底板和打印上去的字进行分离，然后单独作检测以及识别，以提高准确率。

竞赛课题

本竞赛课题正是基于该业务场景提出，考虑到难度我们对这个问题进行了简化和抽象：
一张图片上有两个字符，并有重叠的情况，字符色度差别一般在20-40个色度之间。同一个字符的渲染色度都是一个固定值范围内波动，对重叠部分的色度基本用 alpaha_blending 混合，会增加一定的底板噪声和平滑。
请参赛者请将这两个字符和底板进行分离。语义分割处理后结果以灰度图作为输入：

- 底板：0
- 字符非重叠部分：60 和 120
- 字符重叠部分： 180

字符包括阿拉伯数字，大小写英文和汉字。

评分

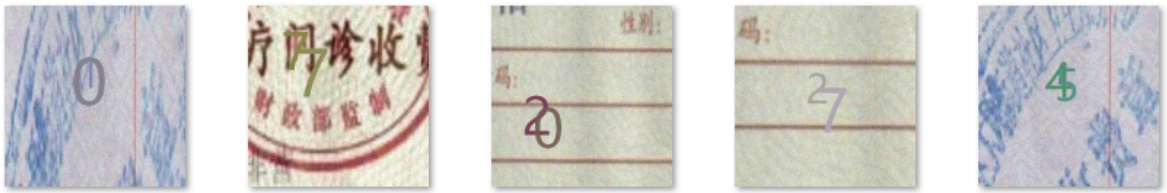
3 月初会发布结果提交页面，包含测试集压缩包，分割结果和代码文档压缩包提交入口，结果提交后不可再更改。
排名分数按照两部分计算：

- 按照格式上传测试集分割结果，平台会计算与测试集样本的交并比，占总分 30%
- 请上传可运行源代码和完整的算法文档压缩包，根据所使用算法的真实场景可推广性，由计算机学院教授和我司资深算法工程师共同打分，占总分 70%

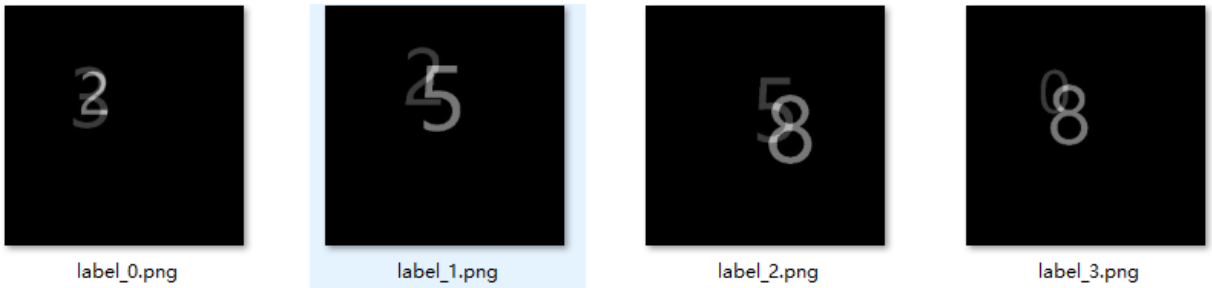
数据集如下所示：

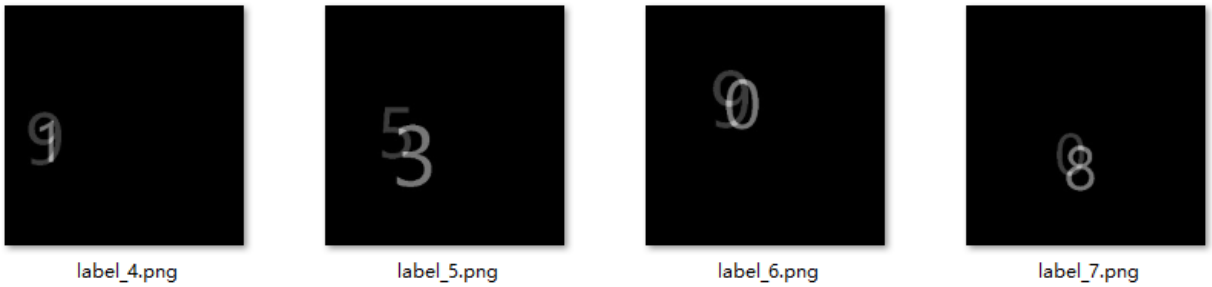

img_0.jpg img_1.jpg img_2.jpg img_3.jpg img_4.jpg


img_5.jpg img_6.jpg img_7.jpg img_8.jpg img_9.jpg


img_10.jpg img_11.jpg img_12.jpg img_13.jpg img_14.jpg

label：


label_0.png label_1.png label_2.png label_3.png


label_4.png label_5.png label_6.png label_7.png

目的就是将两个重叠数字分割开来，从实际场景出发就分割开收据上的重叠文本，以便于后期的信息采集处理。

我们的解决方案

因比赛提供的人工合成数据集和实际场景中的重叠文本差别太大，因此我们的方案也分为两部分：

1. 基于Deeplab网络的文本分割算法

—— 主要针对初赛分割任务

2. 基于弱监督学习的分割识别网络

—— 从实际场景出发解决单据文本重叠方案

方案一_针对初赛的分割任务

在初赛我们主要尝试了三种网络框架：

（1）、具有膨胀卷积的**Deeplab**（线上0.97的指标）

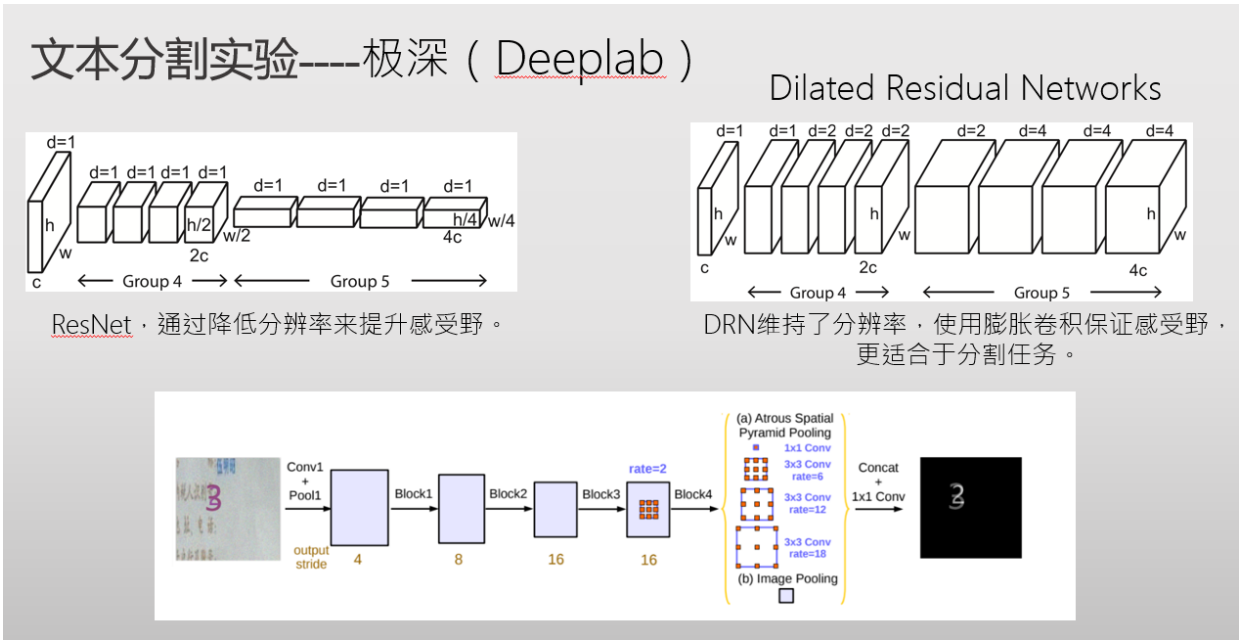
（2）、参考了TextSnake的**FCN**（线上0.955的指标）

（3）、基于VGG11的**U-net**（线上0.965的指标）

Deeplab_v3

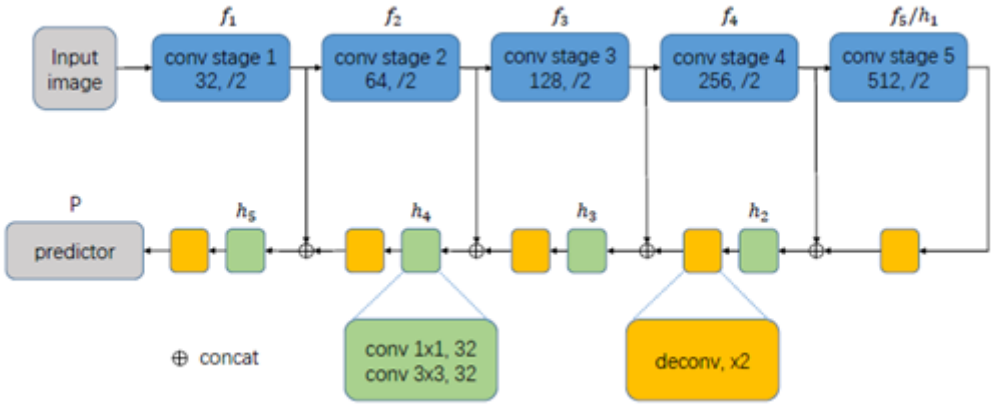
Deeplab网络我就不具体介绍了，网上有很多资料，在分割领域也十分有名

他主要的特殊就是引入了膨胀卷积，提出了**DRN（Dilated Residual Networks）**，他与Resnet，VGG主要的区别如下图



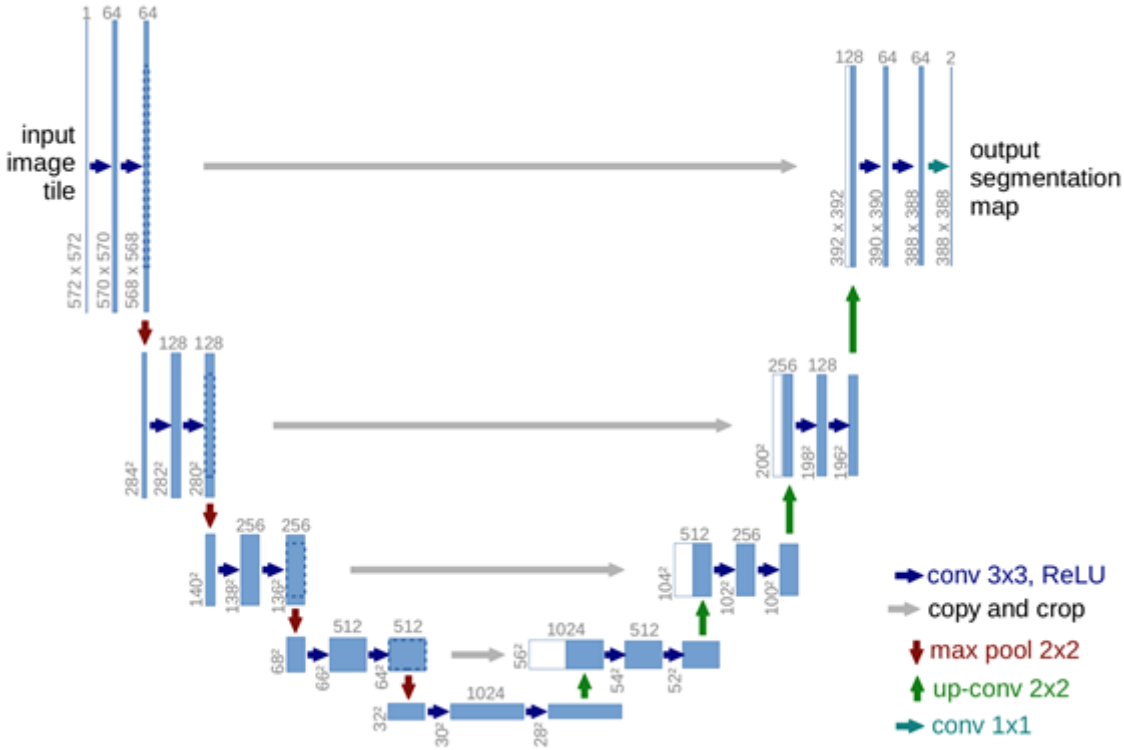
FCN

FCN（Fully Convolutional Network)的一个改进版本作为基本分割网络框架，其实就是TextSnake的网络，基础结构为vgg16，取出下采样中的不同的五层特征进行上采样融合，主要加入的元素有FPN（Feature Pyramid Networks。



FCN

U-net用的是一个kaggle汽车分割冠军方案的网络，基础结构为vgg11。**传送门：**Github



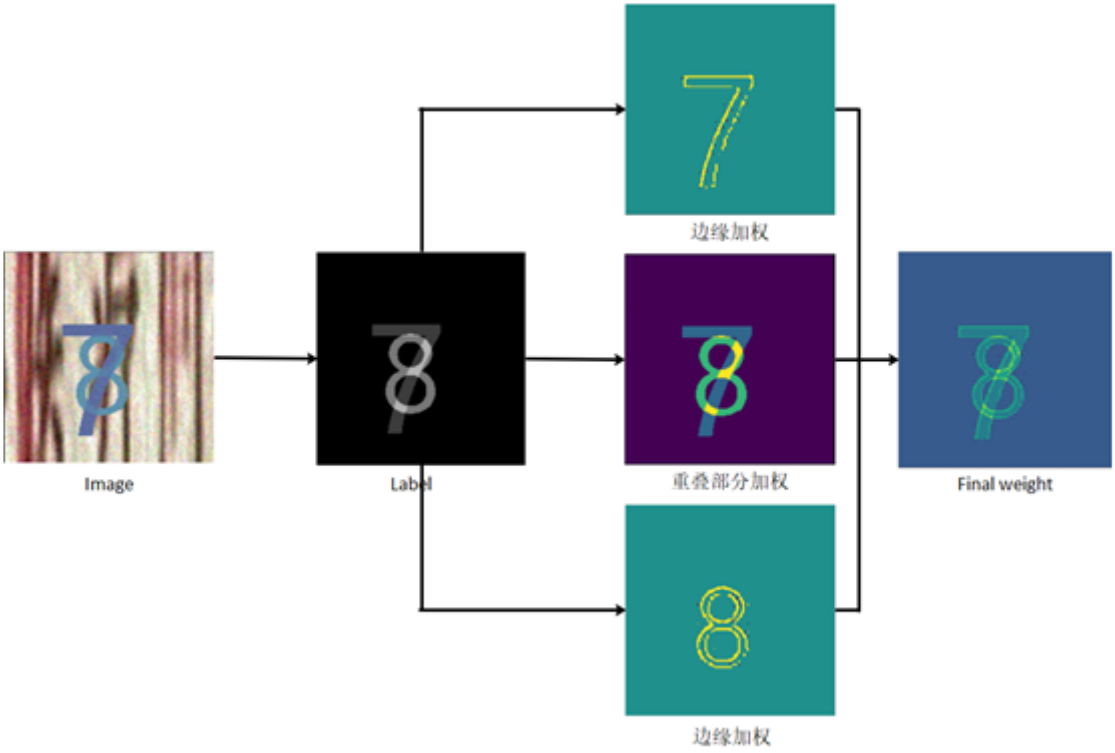
Loss:

用分水岭算法，可以对边缘像素进行加权。

主要是计算 BCE loss 时，mask数字边缘 的像素的权重是在文本数字里面像素的2 倍，而重合部分的像素点是不重合部分像素的2倍.因为考虑到训练到后期，文本内部的像素点的置信度都是很高的，只有在边缘的像素点才有可能被预测错误，而重合部分权重加大是因为分割重叠的数字是本次分割任务的主要目的，并且非重叠区域往往可以分割的较好。

总的Loss函数设计思路：

$$f(x) = BCE + 1 - DICE.$$



方案二_针对现实场景下的弱监督识别网络

刚开始我也尝试用**分割**去分割开打印上去的字体与底板，**发现有如下难点：**

1、现实场景中文本过于精细复杂难以分割。

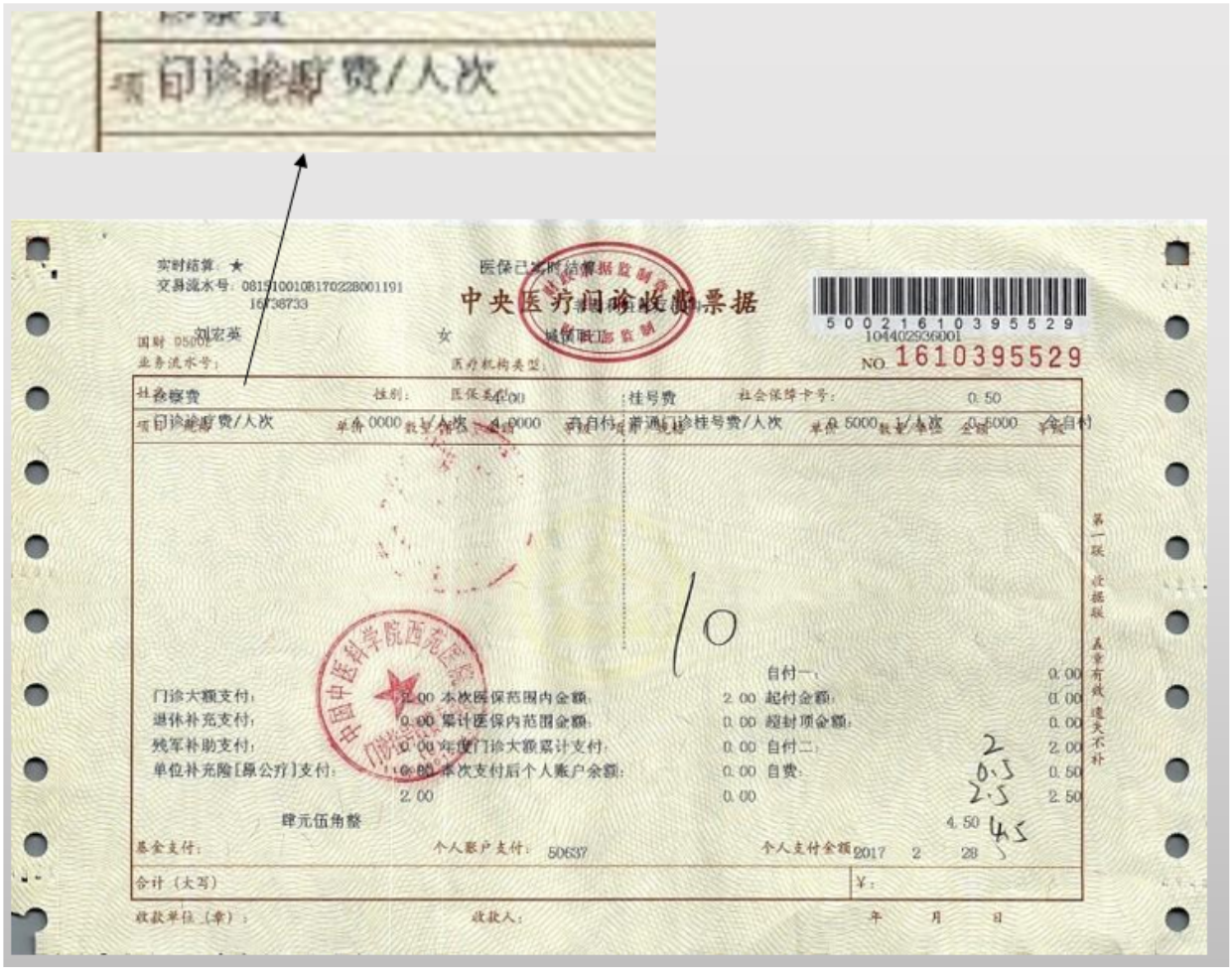
（每个字体的笔画只有2、3像素宽，且干扰严重，这对分割网络要求太高，是很难进行精细分割。）

2、现实场景中的分割label难以获取。

（大多数分割任务的对象都是一个较大的实体对象，而单据中的文字级别的分割label过于精细，基本不可能根据单据上进行标出）

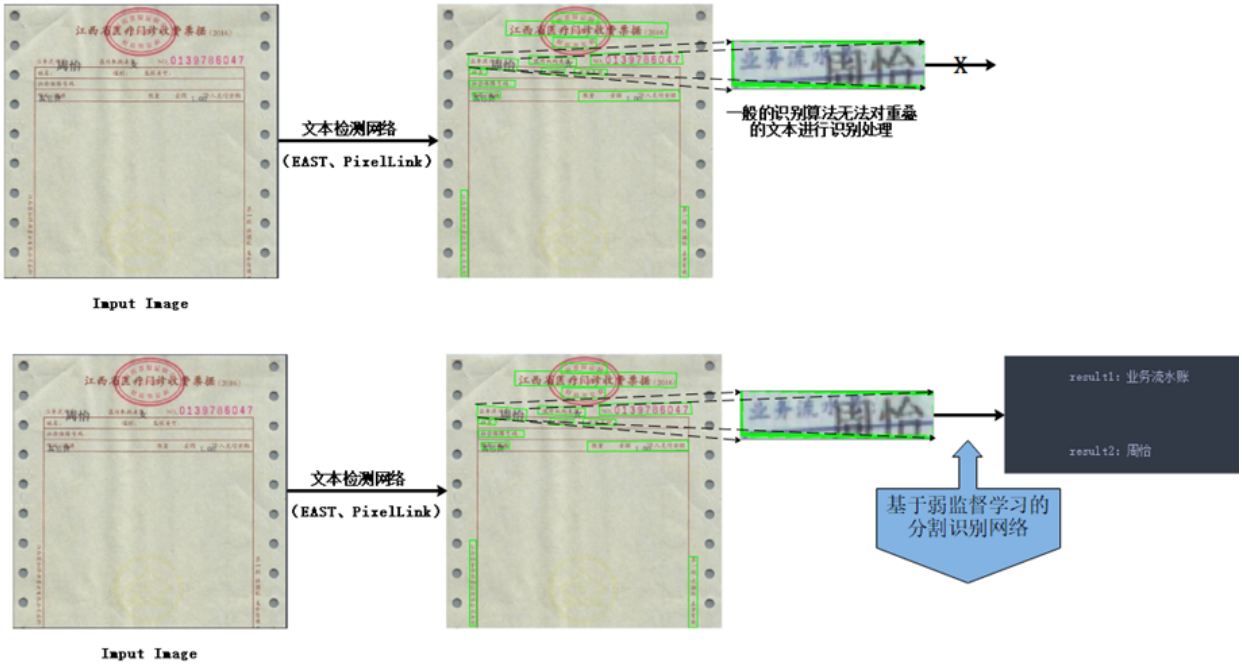
3、现实中的单据形状尺寸相差很大。

其次精细分割任务中忌讳使用resize，而现实中单据形状大小都不一致，甚至可以说差距很大，因此大多数分割网络在单据分割中鲁棒性不强。



我们提出的弱监督分割识别算法到底是做什么事情呢?

弱监督学习分割识别网络 ——— **可以识别重叠文本**

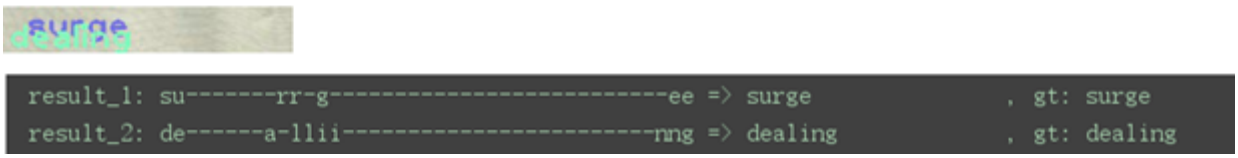


具体算法思路我就不在这里展开了，暂时不能开源。下面是一些人工合成的重叠单词数据集。

数据生成说明：数据底板采样于收据，单词收集于学术论文，颜色随机，**大小仿照真实场合文本大小，**在200*32的图片上随机重合生成（3万训练集，5千测试集）

在该生成数据上，**弱监督分割识别网络baseline在可以达到0.71**的准确率（备注：因时间有限，只用了一天的时间进行的匆忙实验）

准确率 = 预测正确的单词数量/测试集单词总数



反馈与建议

- 微博：@柏林designer
- 邮箱：weijia_wu@yeah.net