挖掘机养成日记

深度学习-OCR_Overview

2019-01-05 | 深度学习, CV | 1 | 2181

8.6k 字 | 31 分钟

本篇涉及使用深度学习的方法实现字符识别的任务，该任务与计算机视觉领域内的图像描述任务（ImageCaption）相似，对于图像描述任务，输入为图像，输出为对该图像的描述；而对于 OCR 任务，输入的图像为包含文字或字符的图片，而输出为这张图片中的文字或字符。

ImageCaption

如图所示，对于 ImageCaption 任务来说，输入一张图片，输出对该图片的描述，简而言之就是看图说话。

OCR

如图所示，对于 OCR 任务，输入的是一张身份张照片，输出的就是该照片中包含的文字内容。

下面结合自己的开发经验对 OCR 项目的一些相关技术思路进行一些记录说明。

技术方案

图像预处理

假如输入时一些扫描或相机拍摄的文档，首先第一点肯定是对图像进行一些预处理，比如灰度处理，二值化处理，仿射变换矫正等对图像进行预处理，使图像统一成设置好的规格。

定位待识别图像

对图像进行预处理之后，接下来就是要找到图像中包含待识别文本的图像，一般分为两种思路，单字符提取和文本行提取

单字符提取

这算是模式识别中比较常用的传统的方法，主要针对的印刷体文档图片，并且英文相比中文的提取效果很好(中文中包含很多左右文字结构，而不像英文都是使用空格进行各个字符分割)。具体的实现方法有如下几种：

- 对图像进行二值化处理，仿射变换等预处理；对图像进行竖直方向投影，提取出文本行；对文本

行在水平方向投影，提取出单个字符，以来设置的阈值进行字符提取，这里就可以看出对于中文来说，进行水平方向投影时，如果阈值设置的不合适，会存在将一个左右结构的汉字拆分成两个单独字符的风险，比如“从”可能会被提取成“人” “人”；

- 假定字符本身是具有连通性的，然后通过连通区域的检测方法找到文字字符的候选。
- 通过最大稳定极值区域（MSER-Maximally Stable Extremal Regions）得到字符的候选，并将这些字符候选看作连通图(graph)的顶点，此时就可以将文本行的寻找过程视为聚类（clustering）的过程，因为来自相同文本行的文本通常具有相同的方向、颜色、字体以及形状。最后使用一个文本分类器滤除非文本部分。
- 北科大殷绪成教授研究组的一个工作对文本的信息进行了更加全面的考虑，使用了文本的颜色、笔画宽度、字符方向（orientation）以及投影的特征。

文本行提取

由于外部因素和内部因素，场景文本检测具有一定的挑战性。外部因素源自环境，例如噪声、模糊和遮挡，它们也是一般目标检测中存在的主要问题。内部因素是由场景文本的属性和变化引起的。与一般目标检测相比，场景文本检测更加复杂，因为：

1. 场景文本可能以任意方向存在于自然图像中，因此边界框可能是旋转的矩形或者四边形；
2. 场景文本边界框的长宽比变化比较大；
3. 因为场景文本的形式可能是字符、单词或者文本行的形式，所以在定位边界的时候算法可能会发生混淆。

基于一般目标检测和语义分割模型，几个精心设计的模型使得文本检测能够更加准确地进行。使用最广泛的是基于 RegionProposal 的方法，其次是基于图像分割的方法。文本检测模型的目标是从图片中尽可能准确地找出文字所在区域。

一般目标检测器（SSD， YOLO 和 DenseBox ）为基础，例如 TextBoxes， FCRN 以及 EAST， SegLink 等，它们直接预测候选的边界框。

以语义分割为基础，例如 PixelLink 和 FTSN，它们生成分割映射，然后通过后处理生成最终的文本边界框。

视觉领域常规物体检测方法(SSD，YOLO，FasterRCNN 等)直接套用于文字检测任务效果并不理想， 主要原因如下：

- ·相比于常规物体，文字行长度、长宽比例变化范围很大。
- ·文本行是有方向性的。常规物体边框BBox的四元组描述方式信息量不充足。
- ·自然场景中某些物体局部图像与字母形状相似，如果不参考图像全局信息将有误报。
- ·有些艺术字体使用了弯曲的文本行，而手写字体变化模式也很多。
- ·由于丰富的背景图像干扰，手工设计特征在自然场景文本识别任务中不够鲁棒。

针对上述问题根因，近年来出现了各种基于深度学习的技术解决方案。它们从特征提取、区域建议网络(RPN)、多目标协同训练、Loss改进、非极大值抑制（NMS）、半监督学习等角度对常规物体检测方法进行改造，极大提升了自然场景图像中文本检测的准确率。例如：

- ·CTPN方案中，用BLSTM模块提取字符所在图像上下文特征，以提高文本块识别精度。
- ·RRPN等方案中，文本框标注采用BBOX +方向角度值的形式，模型中产生出可旋转的文字区域候选框，并
- ·DMPNet等方案中，使用四边形（非矩形）标注文本框，来更紧凑的包围文本区域。
- ·SegLink将单词切割为更易检测的小文字块，再预测邻近连接将小文字块连成词。
- ·TextBoxes等方案中，调整了文字区域参考框的长宽比例，并将特征层卷积核调整为长方形，从而更适
- ·FTSN方案中，作者使用Mask-NMS代替传统BBOX的NMS算法来过滤候选框。
- ·WordSup方案中，采用半监督学习策略，用单词级标注数据来训练字符级文本检测模型。

CTPN (2016年)

论文链接Connectionist Text Proposal Network

该算法由华南理工大学金连文老师研究组提出，该算法脱胎于 FasterRcnn ， CTPN 是目前流传最广、影响最大的开源文本检测模型，可以检测水平或微斜的文本行。在 FasterRcnn 基础上去掉了 ROI 层，引入了不同的 anchor 设置， anchor 设置为固定宽度不同高度，来模拟不同高度的文本行，加入了 RNN 网络，使用 RNN 对目标的位置偏移和置信度得分的计算，具体的实现路径会在接下来进行详细的分析。该算法虽然很准，但是由于使用了 RNN 网络，拖慢了网络速度。

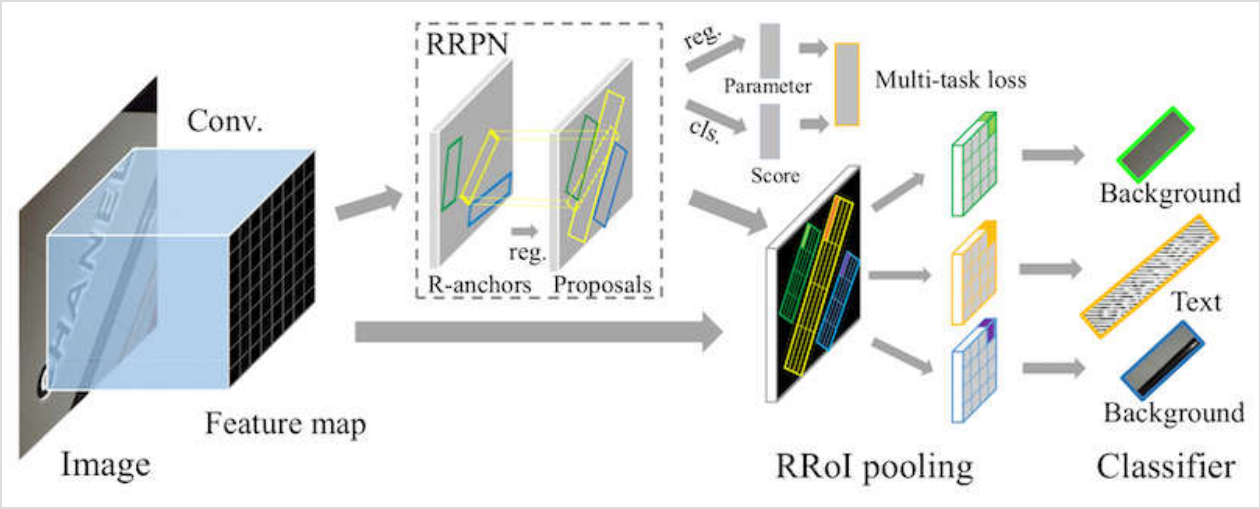
在 CTPN 中文本行可以被看成一个序列 sequence ，而不是一般物体检测中单个独立的目标。同一文本行上各个字符图像间可以互为上下文，在训练阶段让检测模型学习图像中蕴含的这种上下文统计规律，可以使得预测阶段有效提升文本块预测准确率。 CTPN 模型的图像预测流程中，basenet为 VGG16 做基础网络来提取各字符的局部图像特征，之后使用 BLSTM 层提取字符序列上下文特征，并根据 BLSTM 得到的feature为输入进行全连接输出，得到每个anchor的目标预测概率和预测的box，后续的处理部分和fasterRCNN网络的RPN网络相同，经过预测分支输出各个文字块的坐标值和分类结果概率值。在数据后处理阶段，将合并相邻的小文字块为文本行。

RRPN (ECCV2017)

Arbitrary-Oriented Scene Text Detection via Rotation Proposals

这是第一个在场景文字检测中使用RNN的方法，但其主要用于水平文字的场景。基于旋转区域候选网络（RRPN，Rotation Region Proposal Networks）的方案，将旋转因素并入经典区域候选网络（如FasterRCNN）。这种方案中，一个文本区域的ground truth被表示为具有5元组(x, y, h, w, θ)的旋转边框，坐标(x, y)表示边框的几何中心，高度h设定为边框的短边，宽度w为长边，方向是长边的方向。训练时，首先生成含有文本方向角的倾斜候选框，然后在边框回归过程中学习文本方向角。

RRPN中方案中提出了旋转感兴趣区域（RRoI，Rotation Region-of-Interest）池化层，将任意方向的区域建议先划分成子区域，然后对这些子区域分别做max pooling、并将结果投影到具有固定空间尺寸小特征图上。



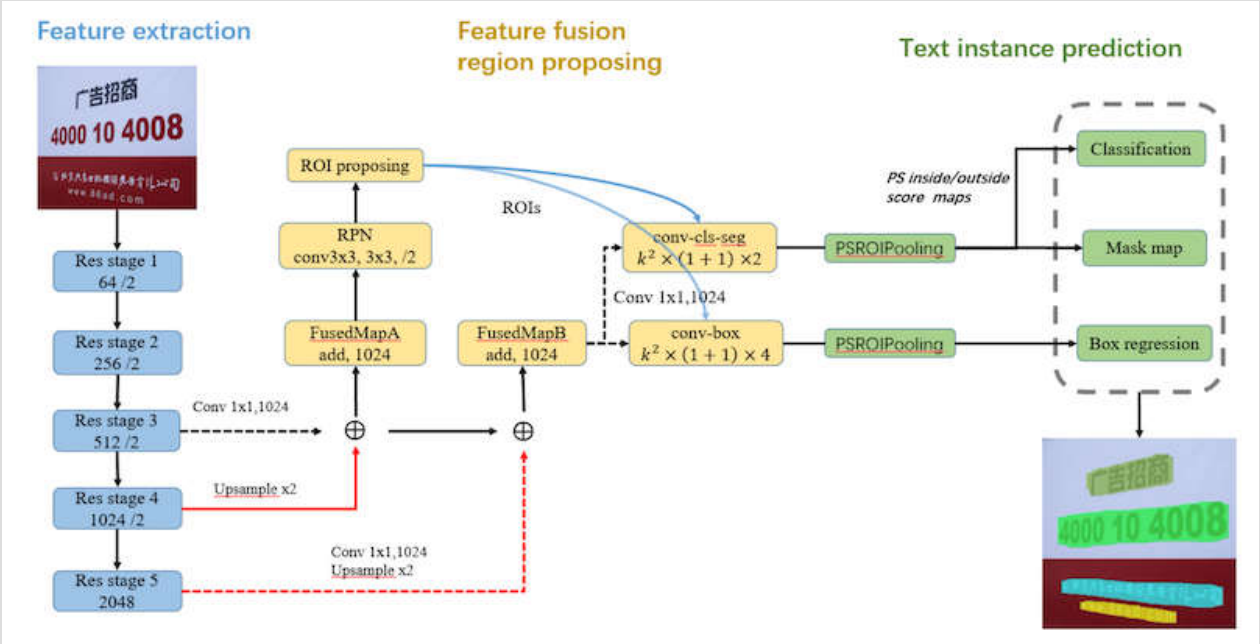
RRPN

FTSN (2018)

Fused Text Segmentation Networks for Multi-oriented Scene Text Detection

FTSN（Fused Text Segmentation Networks）模型使用**分割网络支持倾斜文本检测。**它使用Resnet-101做基础网络，使用了多尺度融合的特征图。标注数据包括文本实例的像素掩码和边框，使用像素预测与边框检测多目标联合训练。

基于文本实例间像素级重合度的Mask-NMS，替代了传统基于水平边框间重合度的NMS算法。下图左边子图是传统NMS算法执行结果，中间白色边框被错误地抑制掉了。下图右边子图是Mask-NMS算法执行结果，三个边框都被成功保留下来。



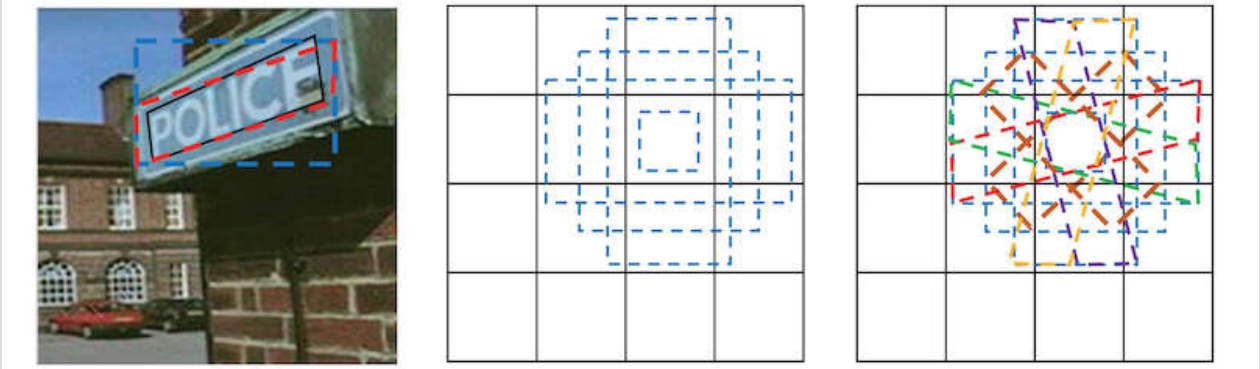
FSTN

`DMPNet` (2017)

[Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection]

`DMPNet` （`Deep Matching Prior Network`）中，使用四边形（非矩形）来更紧凑地标注文本区域边界，其训练出的模型对倾斜文本块检测效果更好。

如下图所示，它使用滑动窗口在特征图上获取文本区域候选框，候选框既有正方形的、也有倾斜四边形的。接着，使用基于像素点采样的 `Monte-Carlo` 方法，来快速计算四边形候选框与标注框间的面积重合度。然后，计算四个顶点坐标到四边形中心点的距离，将它们与标注值相比计算出目标 `Loss` 。文章中推荐用 `LnLoss` 来取代 `L1` 、 `L2Loss` ，从而对大小文本框都有较快的训练回归速度。



(a) Comparison of recalling scene text. (b) Horizontal sliding windows. (c) Proposed quadrilateral sliding windows.

DMPNet_anchors

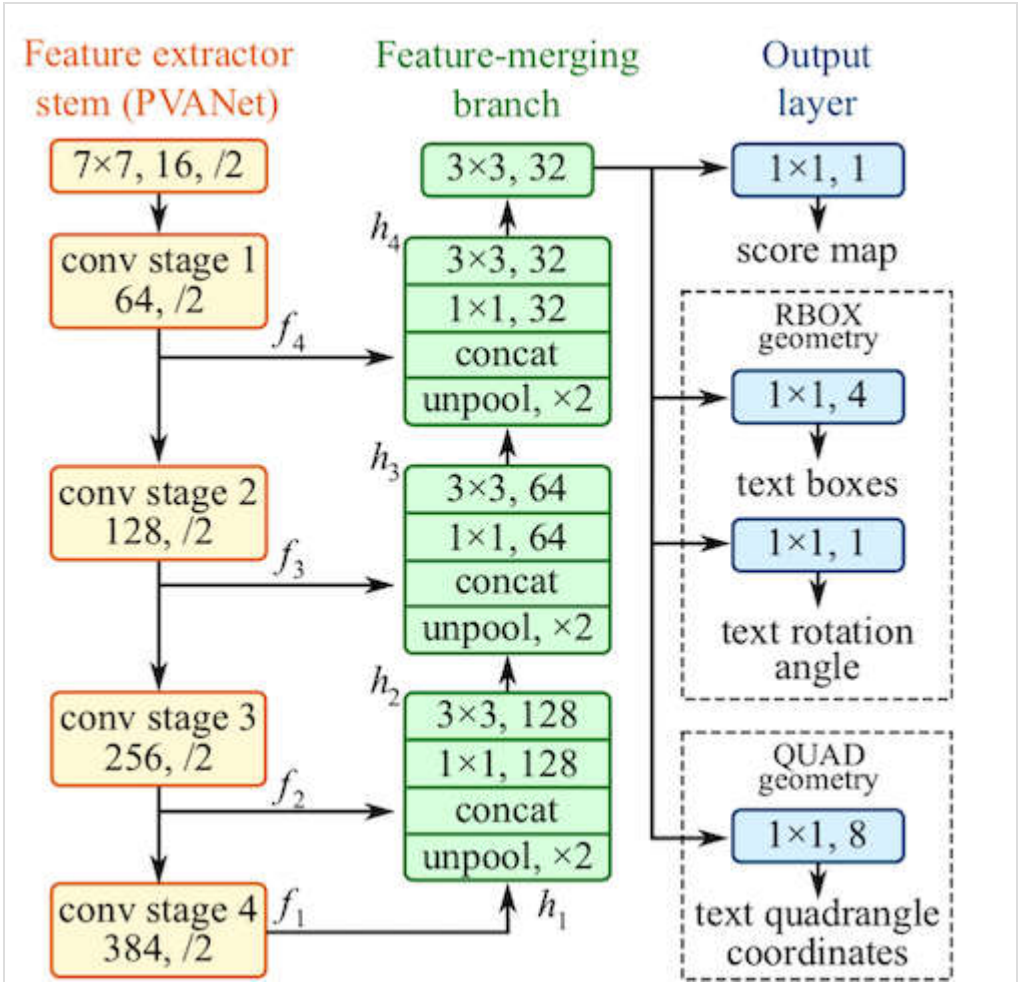
EAST (2017)

An Efficient and Accurate Scene Text Detector

EAST (An Efficient and Accurate Scene Text Detector) 模型中，首先使用全卷积网络（FCN）生成多尺度融合的特征图，作者使用了 PVANet 作为特征提取网络，然后在此基础上直接进行像素级的文本块预测。**该模型中，支持旋转矩形框、任意四边形两种文本区域标注形式。**

- 对应于四边形标注，模型执行时会对特征图中每个像素预测其到四个顶点的坐标差值。
- 对应于旋转矩形框标注，模型执行时会对特征图中每个像素预测其到矩形框四边的距离、以及矩形框的方向角。

上述过程中，省略了其他模型中常见的区域建议、单词分割、子块合并等步骤，因此**该模型的执行速度很快。**



EAST

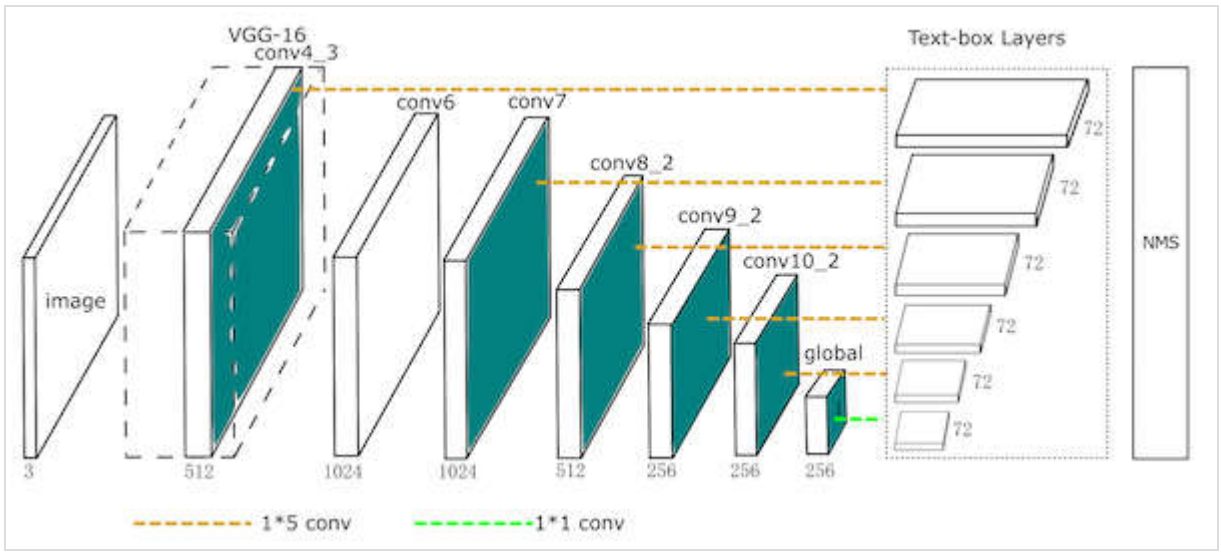
Textboxes(2016)

TextBoxes: A Fast Text Detector with a Single Deep Neural Network

文本检测之TextBoxes

Textboxes 由华中科技大学的白翔组2016年提出。该模型是基于 SSD 框架的图文检测模型，训练方式是端到端的，运行速度也较快。在SSD基础上针对文字的形状做出了如下的改进，模型结构如下图所示

- 为了适应文字行细长型的特点，候选框的长宽比增加了$1, 2, 3, 5, 7, 10$这样初始值；
- 为了适应文本行细长型特点，特征层也用**长条形卷积核**代替了其他模型中常见的正方形$(1 \times 1, 3 \times 3)$卷积核；
- 为了防止漏检文本行，还在垂直方向增加了候选框数量；
- 为了检测大小不同的字符块，在多个尺度的特征图上并行预测文本框， 然后对预测结果做 NMS 过滤；
- 使用识别模型对文字进行过滤和判断，提出了一个实用的 “检测+识别” 的框架。



TextBoxes

SegLink (2017)

论文链接 Detecting Oriented Text in Natural Images by Linking Segments

文本检测之SegLink

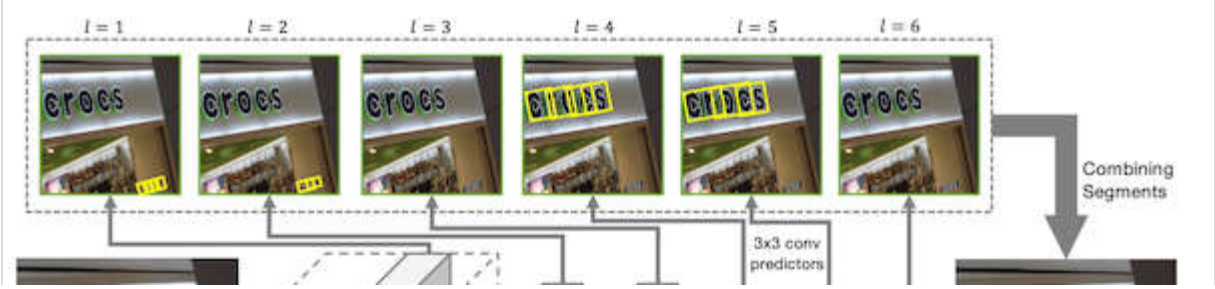
白翔论文分析

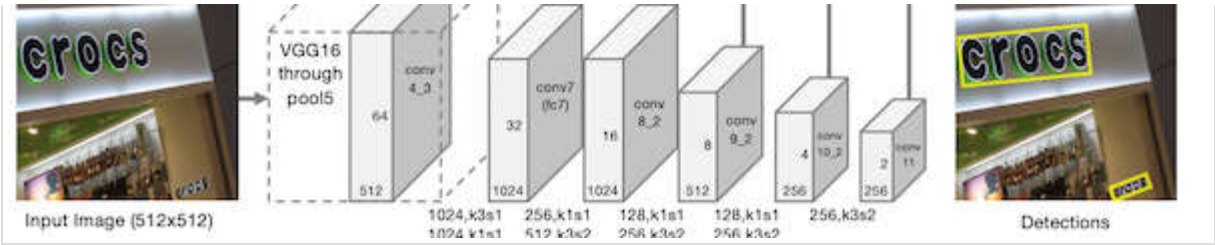
SegLink 模型同样由白翔组提出，由 SSD 改进得来， SegLink 模型的标注数据中，先将每个单词切割为更易检测的有方向的小文字块（$segment = (x, y, w, h, \theta)$），然后用邻近连接（ link ）将各个小文字块连接成单词。这种方案方便于识别长度变化范围很大的、带方向的单词和文本行，它不会像 FasterRCNN 等方案因为候选框长宽比例原因检测不出长文本行。**相比于 CTPN 等文本检测模型， SegLink 的图片处理速度快很多。**

1. 首先给定一个含有文本边界框（ bounding box ）的图片，先使用 CNN 提取图像的特征；
2. 然后用 BLSTM 学习文字的空间上下文信息；
3. 最后对特征进行编码并得到最终的预测结果。

> 与 CTPN 方法相比， SegLink 引入了带方向的 bbox (即文中说的 segment (x, y, w, h, θ))，**它可以检测任意方向的文本行，而 CTPN 主要用于检测水平的文本行，当然如果将垂直 anchor 改成水平 anchor ，也可以检测垂直方向的文本行**；
>
> 与 EAST 方法相比， SegLink 利用不同的 feature map 分别进行预测，而 EAST 是将不同的 feature map 层先进行合并再预测；
>
> 不能检测很大的文本，这是因为 link 主要是用于连接相邻的 segments ，而不能用于检测相距较远的文本行；
>
> 不能检测形变或者曲线文本，这是因为 segments combining 算法在合并的时候采用的是直线拟合．这里可以通过修改合并算法，来检测变形或曲线文本

整个过程可以端到端（ end-to-end ）完成。**提出的定位和识别模型结合之后能得到目前端到端模型中最好的文字检测结果。**





SegLink

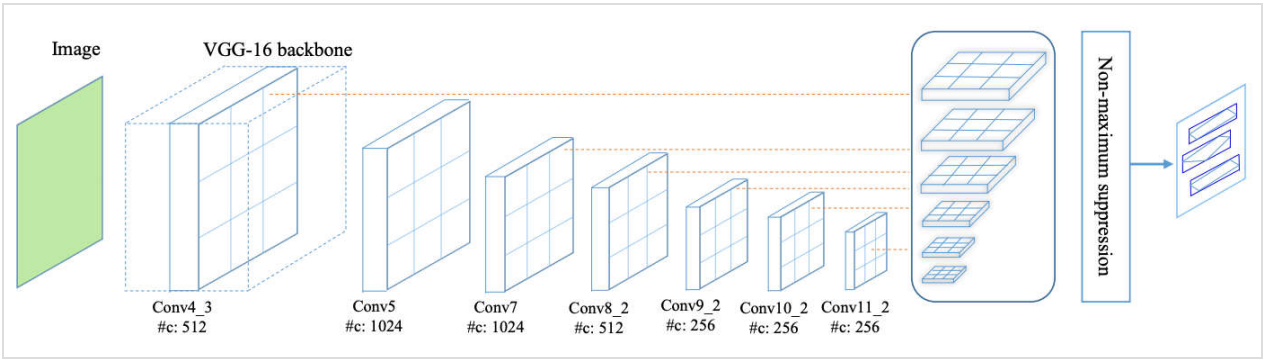
`TextBoxes++` (2018)

[TextBoxes++: A Single-Shot Oriented Scene Text Detector]

[论文笔记：TextBoxes++]

[Github: TextBoxes++]

`Textboxes++` 是 `Textboxes` 的升级版本，目的是增加对倾斜文本的支持。为此，将标注数据改为了旋转矩形框和不规则四边形的格式；对候选框的长宽比例、特征图层卷积核的形状都作了相应调整。·



TextBoxes++

`PixelLink` (2018)

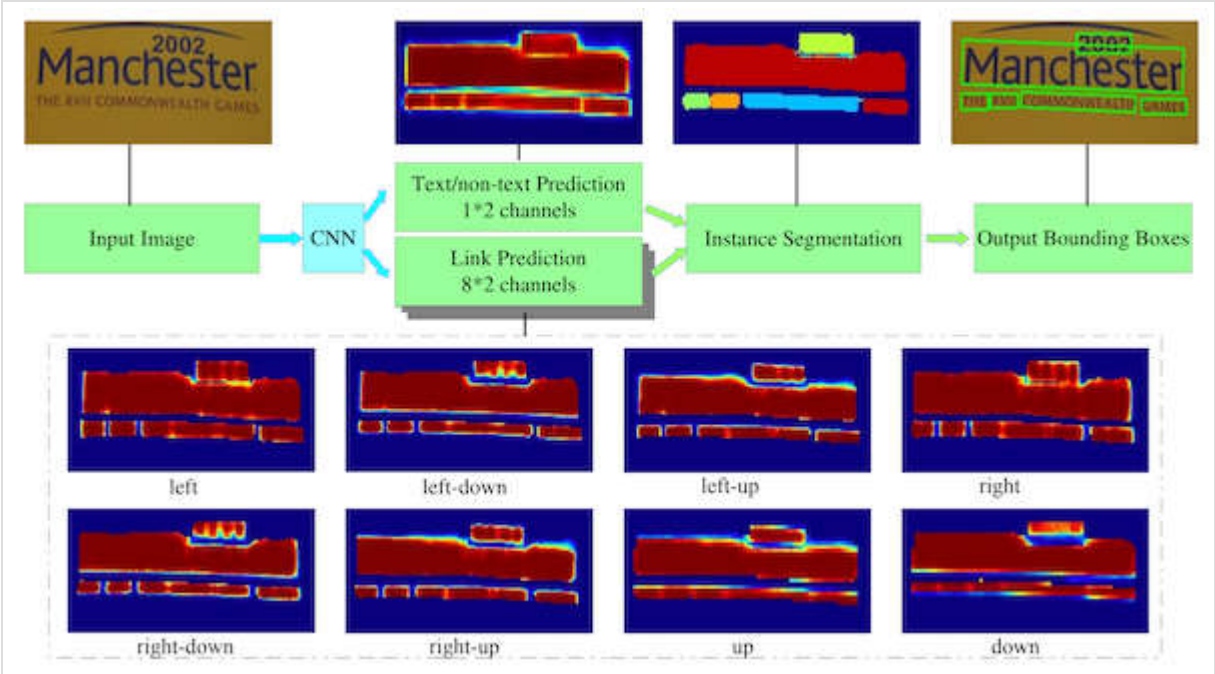
[论文链接 Detecting Scene Text via Instance Segmentation]

自然场景图像中一组文字块经常紧挨在一起，通过语义分割方法很难将它们识别开来，所以阿里开发了 `PixelLink` 模型尝试用**实例分割**方法解决这个问题。

该模型的特征提取部分，为 VGG16 基础上构建的 FCN 网络。模型执行流程如下图所示。首先，借助于 CNN 模块执行两个像素级预测：一个文本二分类预测，一个链接二分类预测。接着，用正链接去连接邻居正文本像素，得到文字块实例分割结果。然后，由分割结果直接就获得文字块边框， 而且允许生成倾斜边框。

上述过程中，**省掉了其他模型中常见的边框回归步骤，因此训练收敛速度更快些。**训练阶段，使用了平衡策略，使得每个文字块在总 LOSS 中的权值相同。训练过程中，**通过预处理增加了各种方向角度的文字块实例。**

- 与CTPN，EAST，SegLink相比，PixelLink放弃了边框回归方法来检测文本行的bbox，而是采用实例分割方法，直接从分割的文本行区域得到文本行的bbox．PixelLink可以以更少额数据和更快地速度进行训练。
- 假设提取特征的主干网络结构采用VGG16(当然也可以采用其它主干网络结构)，PixelLink不需要在imagenet预训练的模型上进行fine-tuned（即直接从头开始训练），而CTPN，EAST，SegLink都需要在imagenet预训练的模型上进行fine-tuned；、
- 与CTPN，EAST，SegLink相比，PixelLink对感受野的要求更少，因为每个神经元值只负责预测自己及其邻域内的状态
- 与SegLink一样，不能检测很大的文本，这是因为link主要是用于连接相邻的segments，而不能用于检测相距较远的文本行



PixelLink

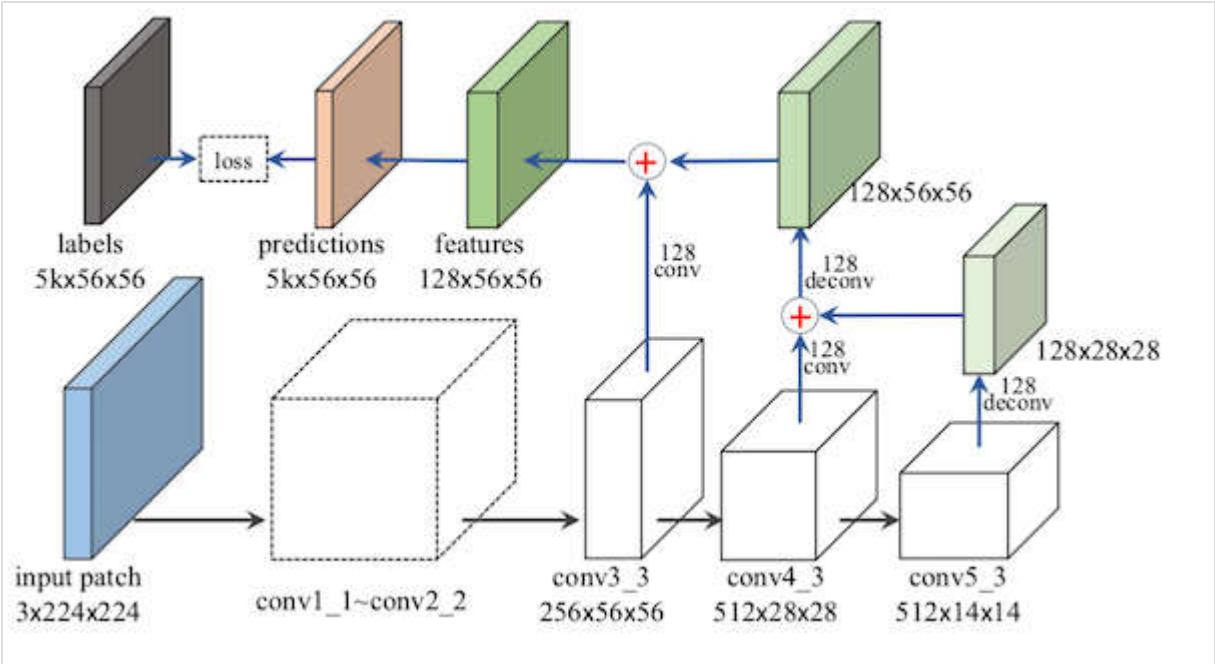
WordSup (2017)

WordSup: Exploiting Word Annotations for Character based Text Detection

百度提出，在数学公式图文识别、不规则形变文本行识别等应用中，字符级检测模型是一个关键基础模块。由于字符级自然场景图文标注成本很高、相关公开数据集稀少，导致现在多数图文检测模型只能在文本行、单词级标注数据上做训练。 WordSup 提出了一种弱监督的训练框架， 可以文本行、单词级标注数据集上训练出字符级检测模型。

WordSup 弱监督训练框架中，两个训练步骤被交替执行：给定当前字符检测模型，并结合单词级标注数据，计算出字符中心点掩码图； 给定字符中心点掩码图，有监督地训练字符级检测模型。

训练好字符检测器后，可以在数据流水线中加入合适的文本结构分析模块，以输出符合应用场景格式要求的文本内容。该文作者例举了多种文本结构分析模块的实现方法。



WordSup

基于角定位与区域分割(2018)

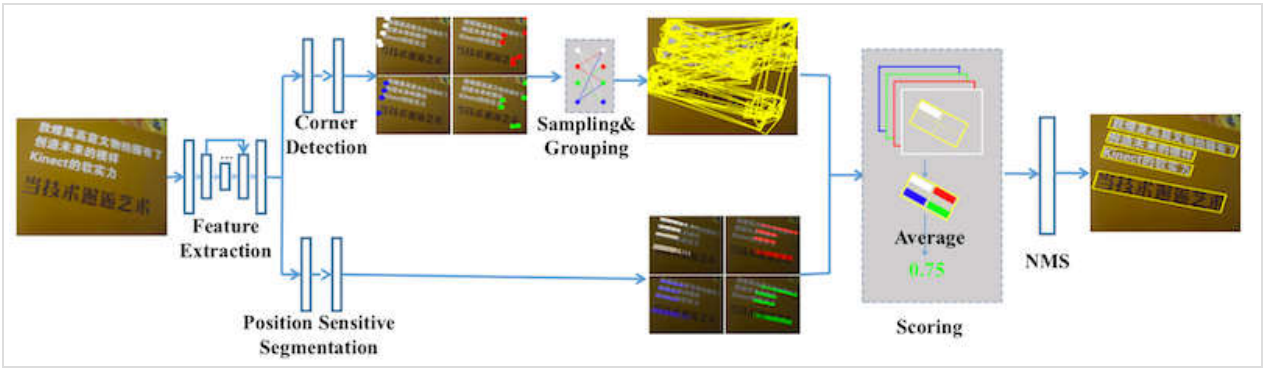
Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation

主要是为了解决场景文本多方向，长宽比变化较大等场景文本检测中的难点问题。

该方法用一个端到端网络完成文字检测整个过程——除了基础卷积网络（ backbone ）外，包括两个并行分支和一个后处理。第一个分支是通过一个DSSD网络进行角点检测来提取候选文字区域，第二个分支是利用类似于 RFCN 进行网格划分的方式来做 position-sensitive 的 segmentation 。后处理是利用 segmentation 的 score map 的综合得分，过滤角点检测得到的候选区域中的噪声。

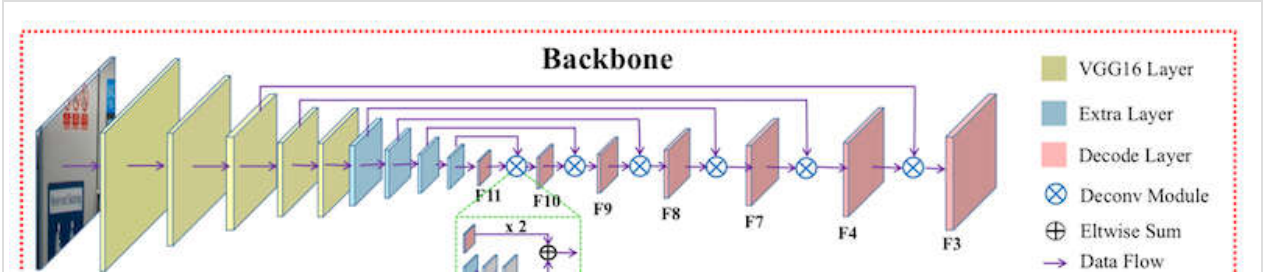
文章亮点

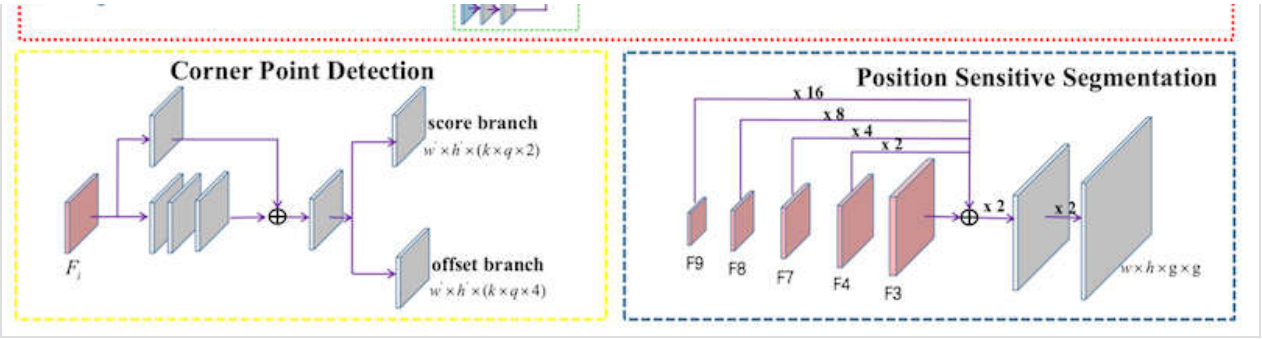
- 检测不是用一般的 object detection 的框架来做，而是用 corner point detection 来做。（可以更好解决文字方向任意、文字长宽比很大的文本，不会受到感受野的影响）
- 分割用的是 position sensitive segmentation ，仿照 RFCN 划分网格的思路，把位置信息融合进去（对于检测单词这种细粒度的更有利）
- 把检测+分割两大类的方法整合起来，进行综合打分的 pipeline （可以使得检测精度更高）



Overview of our method. Given an image, the network outputs corner points and segmentation maps by corner detection and position-sensitive segmentation. Then candidate boxes are generated by sampling and grouping corner points. Finally, those candidate boxes are scored by segmentation maps and suppressed by NMS

- **backbone**：**基础网络，用来特征提取（不同分支特征共享）**
- **corner detection**：用来生成候选检测框，是一个独立的检测模块，类似于RPN的功能
- **Position Sensitive Segmentation**：整张图逐像素的打分，和一般分割不同的是输出4个score map，分别对应左上、左下、右上、右下不同位置的得分
- **Scoring + NMS**：综合打分，利用（2）的框和（3）的score map再综合打分，去掉非文字框，最后再接一个NMS



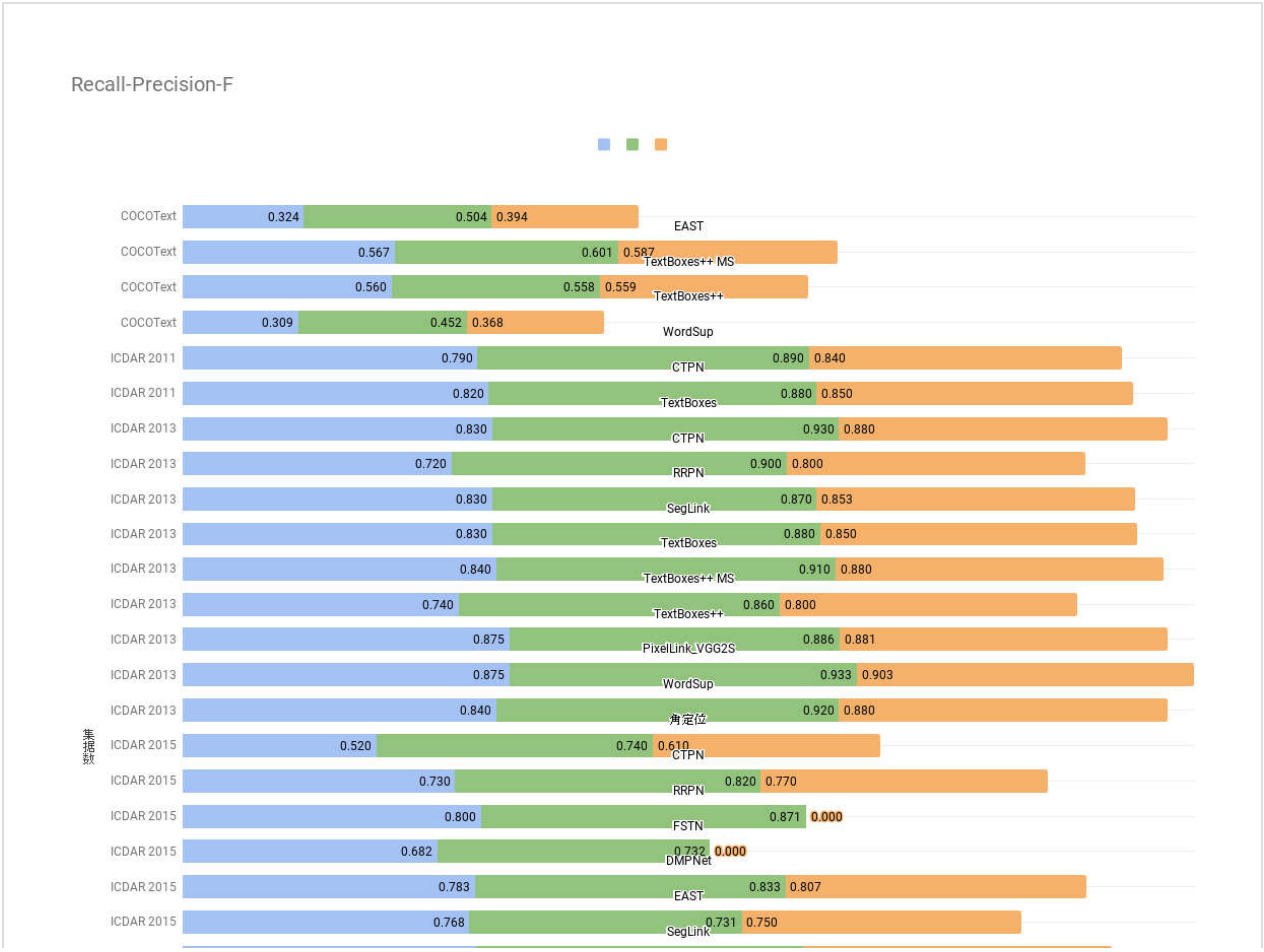


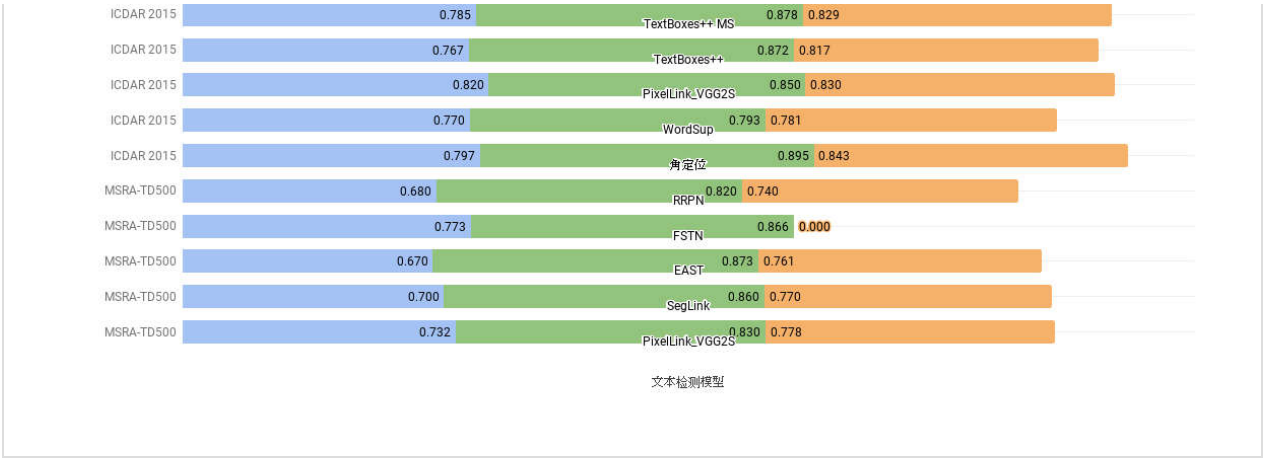
Network Architecture

- Backbone 取自DSSD = VGG16(pool5) + conv6(fc6) + conv7(fc7) + 4conv + 6 deconv (with 6 residual block)
- Corner Point Detection 是类似于 SSD ， 从多个deconv的feature map上单独做detection得到候选框，然后多层的检测结果串起来nms后为最后的结果

总结

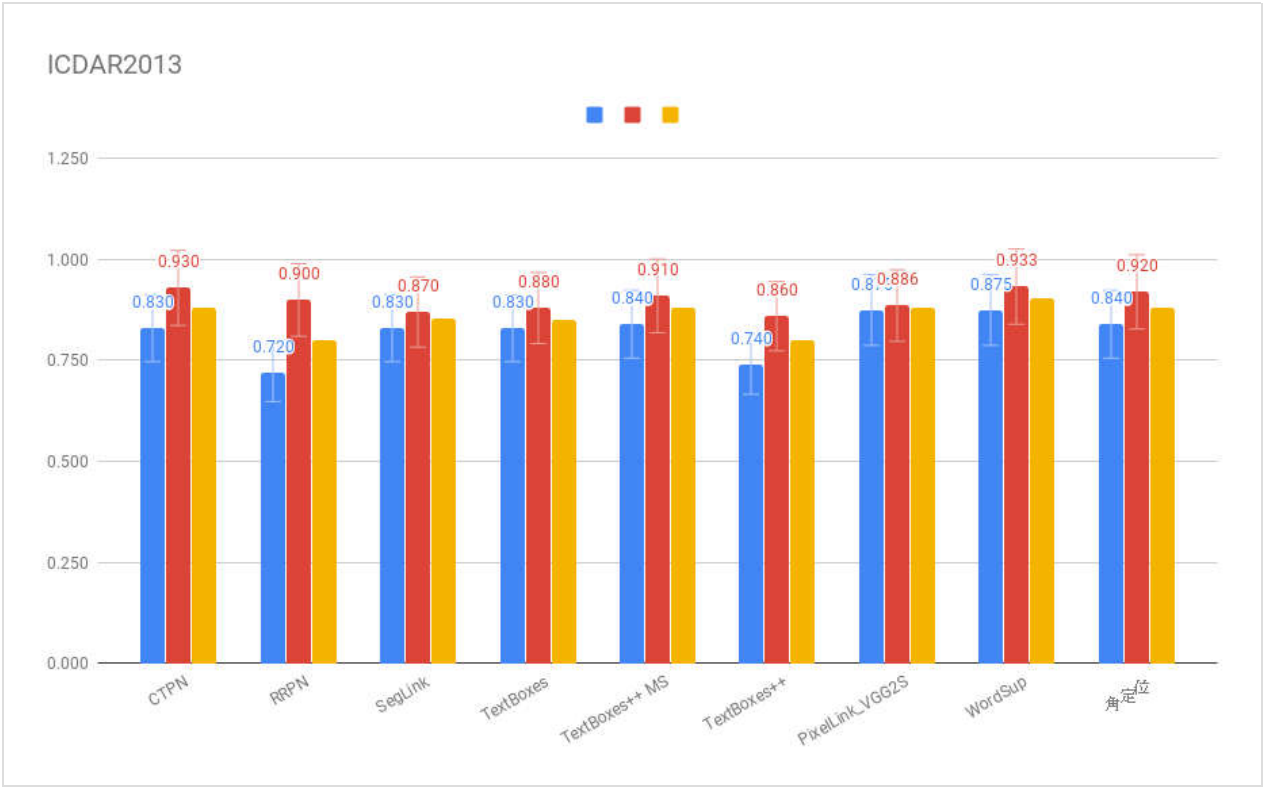
- 多个数据集中不同模型的表现





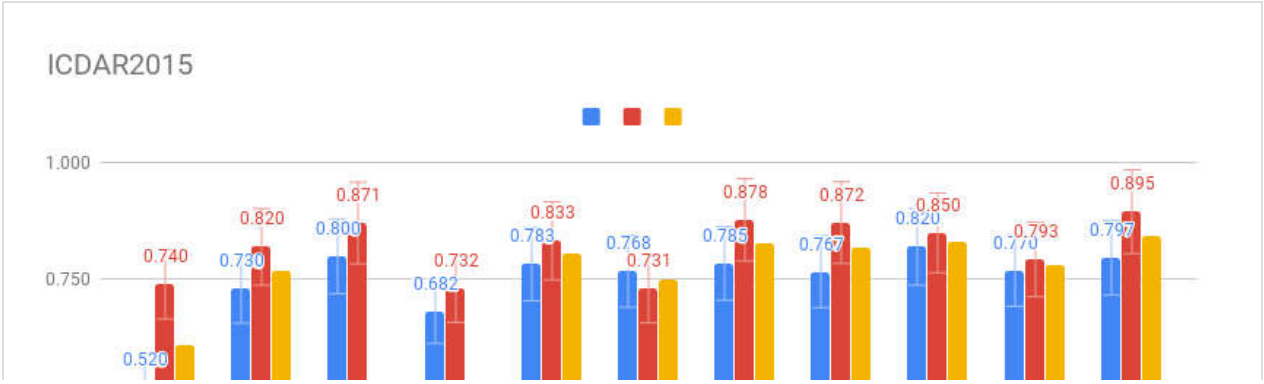
overview

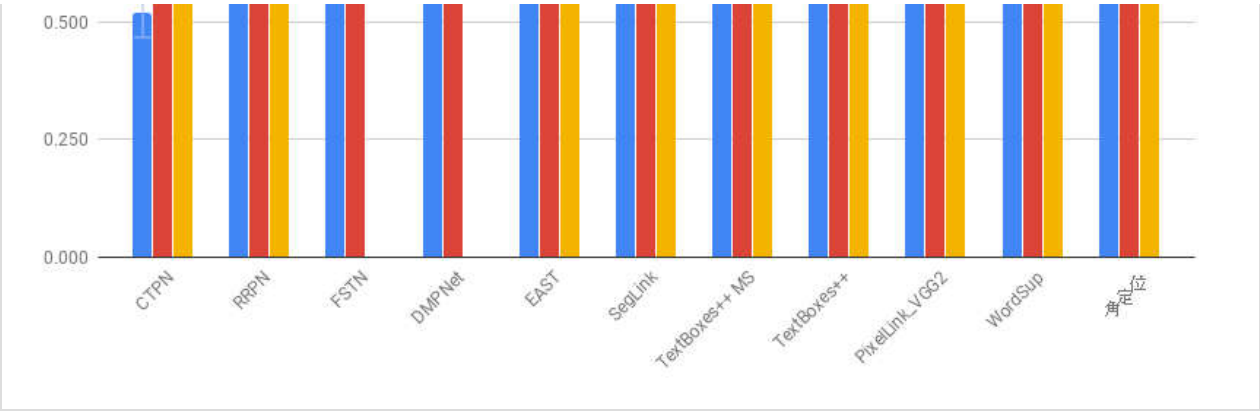
- ICDAR2013数据集表现



ICDAR2013

- ICDAR2015数据集表现





ICDAR2015

图中蓝色为召回率(实际为真的样本中预测为真的概率，衡量检索的覆盖率)，绿色为精确率(预测为真的样本中实际为真的概率，衡量检测的信噪比)，橙色为Fscore，图中显示0.00的数据表示论文中没有给出相应得分。具体可以参考另一篇文章机器学习-常用知识点[可以直观的看到各个模型的评价指标值，在IC2015数据集上， `PixelLink_vgg2` 对应的模型效果是最好的

文本识别

提取待识别图像之后就可以对图像进行识别了，和定位待识别图像相对应，同样是两种类型识别任务

单字符识别

k近邻

该方法很好理解，数据库中保存着已经标注好label的不同的字符的图像，对于新的图像，识别时，比较待识别的图像和数据库中图像的距离，选择距离最小的类别对应的label为当前图像的识别结果，但是该方法存在一个问题，那就是计算量太大了，要想提高识别的准确性，就需要数据库中存储的数据要足够多，与此同时，在识别时计算量也会增加，识别一个图像要跟整个数据库中的所有样本进行距离计算

图像分类模型

借着深度学习的东风，使用图像识别算法进行单字符识别效果简直不要太好，相关的技术方法有很多，包括 `VggNet` ， `ResNet` ， `InceptionNet` ， `DenseNet` 等，具体的可以参考本人的另一篇文字深度学习-图像识别。

不定长字符识别

目前基本都是使用的 `CTC` 策略(稍后详解)，对于特征提取的基础网络目前主要使用的是 `CRNN` 和

Densenet

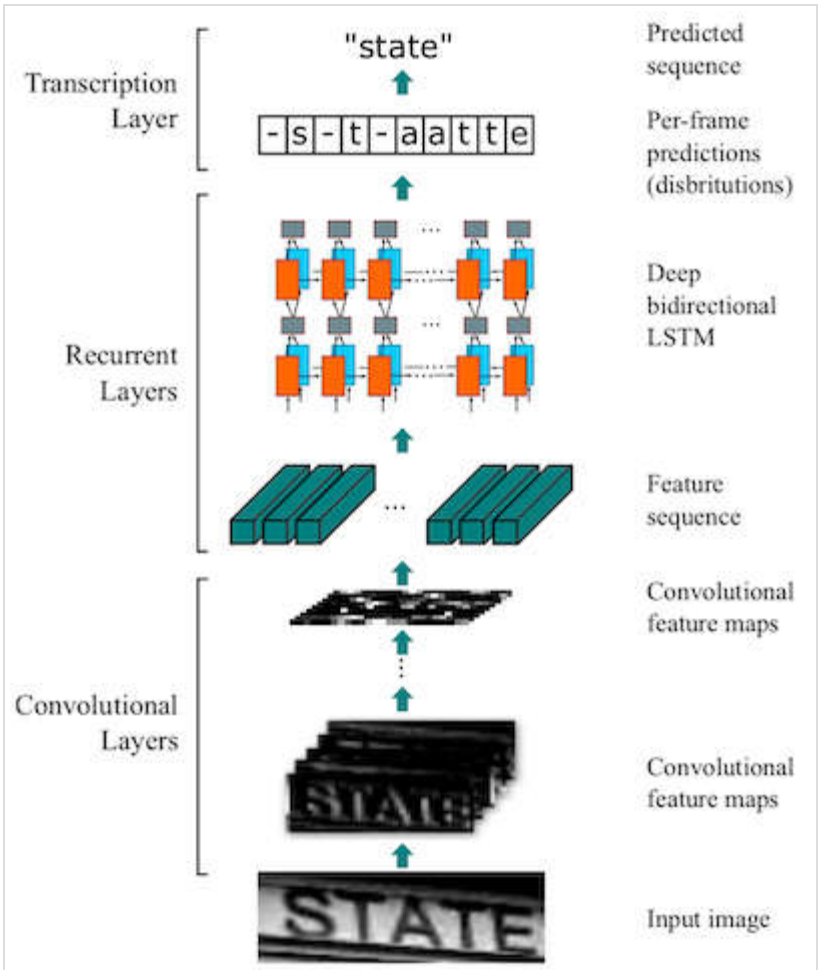
CRNN+CTC (2015)

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

CRNN（Convolutional Recurrent Neural Network）是目前较为流行的图文识别模型，可识别较长的文本序列。

- 包含 CNN 特征提取层和 BLSTM 序列特征提取层，能够进行端到端的联合训练。
- 利用 BLSTM 和 CTC 部件学习字符图像中的上下文关系， 从而有效提升文本识别准确率，使得模型更加鲁棒。

预测过程中，前端使用标准的 CNN 网络提取文本图像的特征，利用 BLSTM 将特征向量进行融合以提取字符序列的上下文特征，然后得到每列特征的概率分布，最后通过转录层(CTC rule)进行预测得到文本序列。



CRNN

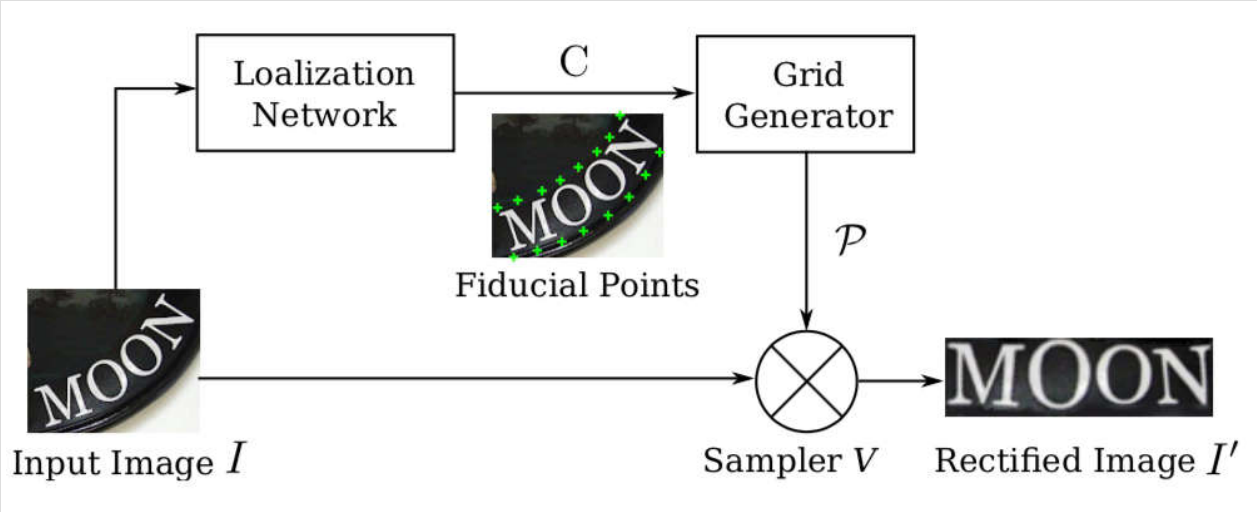
DenseNet+CTC

该框架和 CRNN 相似，使用 DenseNet 替换了原来的 CNN+BLSM ，提升了模型的运行速度，但是DenseNet由于block层对特征图的重复利用会非常占显存。

RARE (2016)

Robust Scene Text Recognition with Automatic Rectification

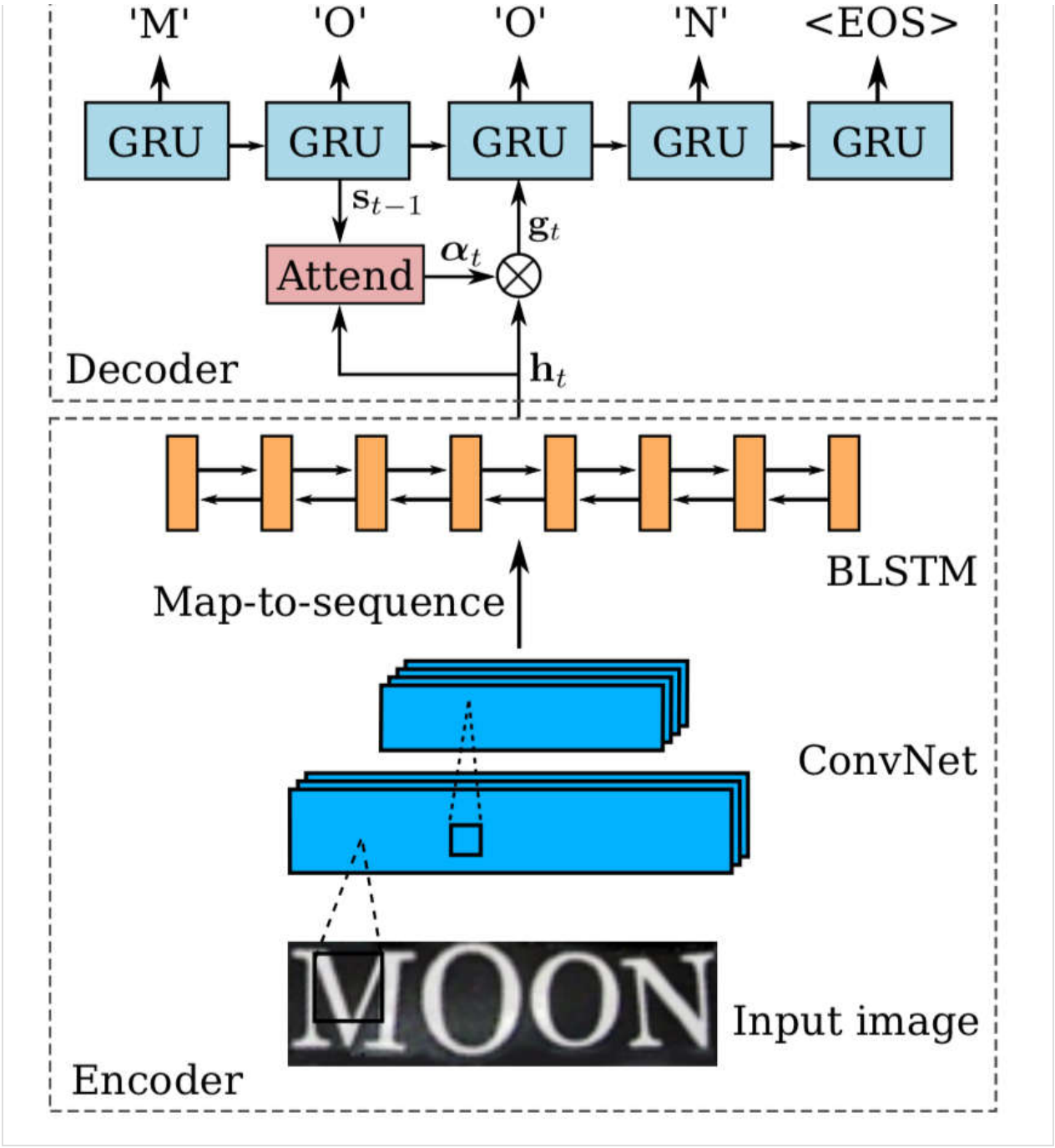
RARE 模型在识别变形的图像文本时效果很好。模型预测过程中，输入图像首先要被送到一个空间变换网络(Spatial Transformer Network)中做处理，矫正过的图像然后被送入序列识别网络中得到文本预测结果。空间变换网络如下图所示。网络模型有 STN+SRN 组成。



STNet

空间变换网络内部包含定位网络(Localization Network)、网格生成器(Grid Generator)、采样器(Sampler)三个部件。经过训练后，它可以根据输入图像的特征图动态地产生空间变换网格，然后采样器根据变换网格核函数从原始图像中采样获得一个矩形的文本图像。 RARE 中支持一种称为 TPS (thin-plate splines) 的空间变换，从而能够比较准确地识别透视变换过的文本、以及弯曲的文本。

经过矫正之后的图像送入序列识别网络(Sequence Recognition Network)，网络结构如下图所示，为一个 Encoder_Decoder 模型，结构为 CNN+BLSTM+ATT_GRU+Softmax ，使用极小化对数似然估计的方法，优化方法为 ADADELTA



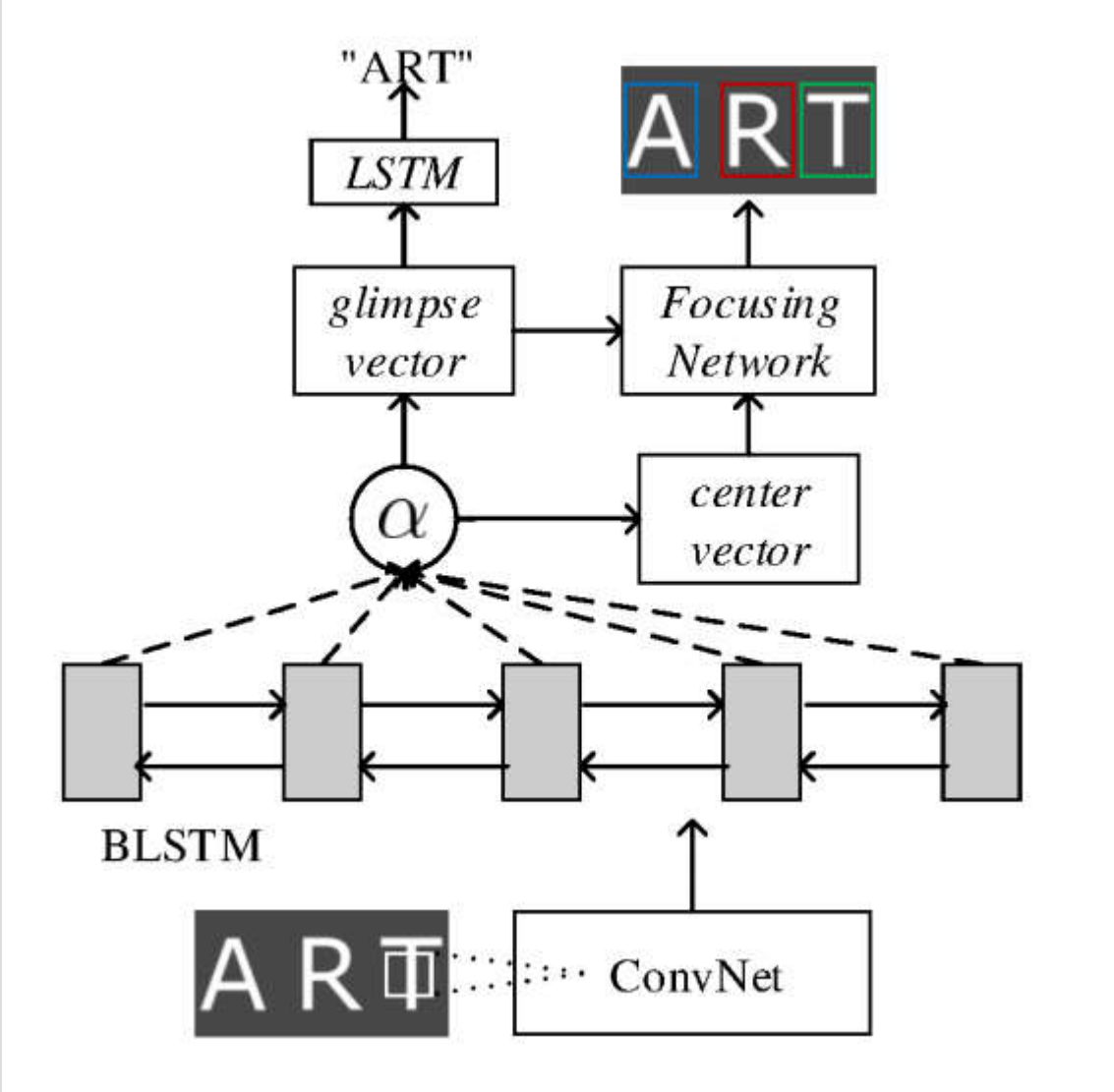
SRN

`FocusingAttention` (2017)

[Focusing Attention: Towards Accurate Text Recognition in Natural Images]

利用 `attention model` 去做序列文字识别，可能会因为图像分辨率较低、遮挡、文字间间隔较大等问题而导致 `attention` 位置并不是很准，从而造成字符的错误识别。海康威视在ICCV2017上提出使用字

符像素级别的监督信息使 attention 更加准确地聚焦在文字区域，从而使识别变得更精准。他们用了部分像素级别的标注，有了类别信息以后做多任务，结果较为精准。并且只要部分字符的标注就可以带来网络性能的一定提升。



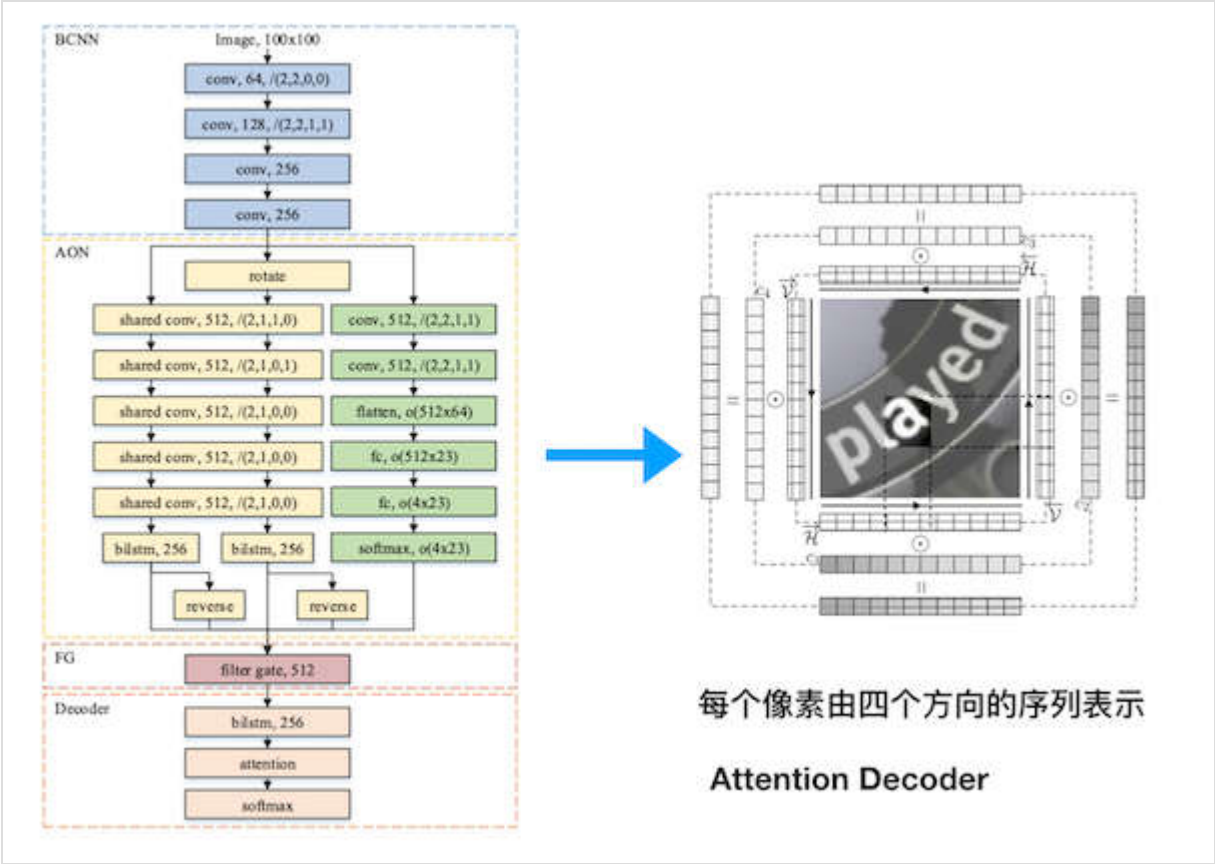
FocusingAttention

AON (2018)

AON: Towards Arbitrarily-Oriented Text Recognition

针对有形变或者任意方向文字的识别问题，Cheng等人在CVPR2018上提出了该模型。他们在水平方向之外加了一个竖直方向的双向LSTM，这样的话就有从上到下，从下到上，从左到右，从右到左四个方向序列的特征建模。接下来引入一个权重，该权重用来表示来自不同方向的特征在识别任务中发挥作

用的重要性。这对性能有一定提升，尤其是对任意排列的文字识别。



AON

ASTER(2018)

ASTER_An Attentional Scene Text Recognizer with Flexible Rectification

主要解决**不规则排列文字**的**文字识别**问题，论文为之前一篇 RARE 的改进版（journal版）

主要思路:

- 针对不规则文字，先矫正成正常线性排列的文字，再识别；
- 整合矫正网络和识别网络成为一个端到端网络来训练；
- 矫正网络使用STN，识别网络用经典的 `seq2seq+attention`

和 STN 的不同点

本文在输入网络前将原图resize成小的图，然后在该小图上预测control point，而输入到Grid Generator或Sample计算的时候又映射回原图大小。这样的目的是**为了减小网络参数，降低计算量**（但有没有可能小图对于control point的prediction会不准？对于识别来讲，每个word的patch块本

身就比较小了，而且小图映射回大图的点位置这个误差比例就会放大？）

和 RARE 的不同点

网络最后fc层的激活函数不是用tanh，而是直接对值进行clipping（具体怎么clip论文没说），这样做的目的是**为了解决采样点可能落到图外面的问题，以及加快了网络训练的收敛速度**，论文中对此没有解释本质原因，只是说明实验证明如此。

按照论文所述，ASTER和RARE之间的网络结构基本没有大幅度的变化，仍然是先矫正后识别，并且矫正模型和识别模型基本相同，但是效果相比RARE好了太多，不知为何。ASTER的网络参数配置如下所示

	Layers	Out Size	Configurations
Encoder	Block 0	32×100	3×3 conv, s 1×1
	Block 1	16×50	$\begin{bmatrix} 1 \times 1 \text{ conv}, 32 \\ 3 \times 3 \text{ conv}, 32 \end{bmatrix} \times 3$, s 2×2
	Block 2	8×25	$\begin{bmatrix} 1 \times 1 \text{ conv}, 64 \\ 3 \times 3 \text{ conv}, 64 \end{bmatrix} \times 4$, s 2×2
	Block 3	4×25	$\begin{bmatrix} 1 \times 1 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 128 \end{bmatrix} \times 6$, s 2×1
	Block 4	2×25	$\begin{bmatrix} 1 \times 1 \text{ conv}, 256 \\ 3 \times 3 \text{ conv}, 256 \end{bmatrix} \times 6$, s 2×1
	Block 5	1×25	$\begin{bmatrix} 1 \times 1 \text{ conv}, 512 \\ 3 \times 3 \text{ conv}, 512 \end{bmatrix} \times 3$, s 2×1
	BiLSTM 1	25	256 hidden units
	BiLSTM 2	25	256 hidden units
Decoder	Att. LSTM	*	256 attention units 256 hidden units
	Att. LSTM	*	256 attention units 256 hidden units

ASTER网络参数

总结

目前主要的 OCR 项目使用的是如上的这些技术，比如 CTPN+CRNN， CTPN+DenseNet， YOLO+DenseNet， PixelLink+RARE 等，接下来针对上面这些技术点结合论文及代码进行一些个人的浅薄分析。

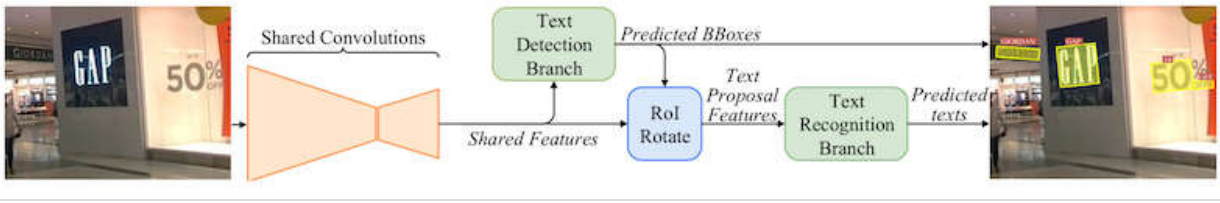
End2End

端对端模型，直接从图片中定位并识别出包含的文本内容

FOTS (2018)

FOTS: Fast Oriented Text Spotting with a Unified Network

FOTS 是图像文本检测与识别同步训练、端到端可学习的网络模型。检测和识别任务共享卷积特征层，既节省了计算时间，也比两阶段训练方式学习到更多图像特征。引入了旋转感兴趣区域（RoIRotate），可以从卷积特征图中产生出定向的文本区域，从而支持倾斜文本的识别。

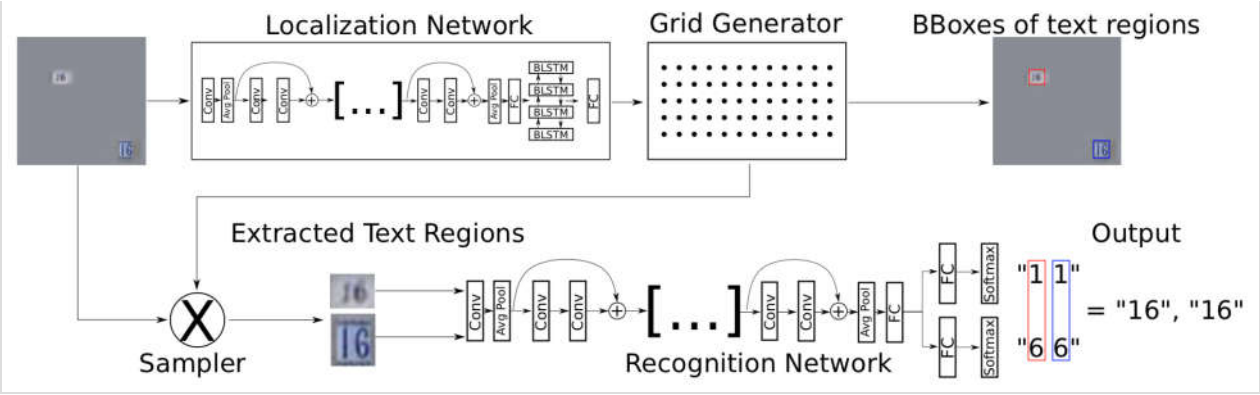


FOTS

STN-OCR (2017)

STN-OCR: A single Neural Network for Text Detection and Text Recognition

STN-OCR 是集成了了图文检测和识别功能的端到端可学习模型。在它的检测部分嵌入了一个空间变换网络（STN）来对原始输入图像进行仿射（affine）变换，类似于 RARE 中的 STN 网络。利用这个空间变换网络，可以对检测到的多个文本块分别执行旋转、缩放和倾斜等图形矫正动作，从而在后续文本识别阶段得到更好的识别精度。在训练上 STN-OCR 属于**半监督学习方法，只需要提供文本内容标注，而不要求文本定位信息。**作者也提到，**如果从头开始训练则网络收敛速度较慢，因此建议渐进地增加训练难度。** STN-OCR 已经开放了工程源代码和预训练模型。



STN-OCR

Mask TextSpotter(2018)

该文受到 MaskRCNN 的启发提出了一种用于场景text spotting的可端到端训练的神经网络模型：M ask TextSpotter 。与以前使用端到端可训练深度神经网络完成text spotting的方法不同，Mask TextSpotter利用简单且平滑的端到端学习过程，通过语义分割获得精确的文本检测和识别。此外，它在处理不规则形状的文本实例（例如，弯曲文本）方面优于之前的方法。

数据集

训练数据集

本章将列举可用于文本检测和识别领域模型训练的一些大型公开数据集， 不涉及仅用于模型fine-tune任务的小型数据集。

Chinese Text in the Wild(CTW)

该数据集包含32285张图像，1018402个中文字符(来自于腾讯街景), 包含平面文本，凸起文本，城市文本，农村文本，低亮度文本，远处文本，部分遮挡文本。图像大小2048*2048，数据集大小为31GB。以(8:1:1)的比例将数据集分为训练集(25887张图像，812872个汉字)，测试集(3269张图像，103519个汉字)，验证集(3129张图像，103519个汉字)。

```
1  文献链接：https://arxiv.org/pdf/1803.00085.pdf
2  数据集下载地址：https://ctwdataset.github.io/
```

Reading Chinese Text in the Wild(RCTW-17)

该数据集包含12263张图像，训练集8034张，测试集4229张，共11.4GB。大部分图像由手机相机拍摄，含有少量的屏幕截图，图像中包含中文文本与少量英文文本。图像分辨率大小不等。

```
1  下载地址http://mclab.eic.hust.edu.cn/icdar2017chinese/dataset.html
2  文献：http://arxiv.org/pdf/1708.09585v2
```

ICPR MWI 2018 挑战赛

大赛提供20000张图像作为数据集，其中50%作为训练集，50%作为测试集。主要由合成图像，产品描述，网络广告构成。该数据集数据量充分，中英文混合，涵盖数十种字体，字体大小不一，多种版式，背景复杂。文件大小为2GB。

```
1  下载地址：
2  https://tianchi.aliyun.com/competition/information.htm?raceId=231651&_is_login_redi
```

Total-Text

该数据集共1555张图像，11459文本行，包含水平文本，倾斜文本，弯曲文本。文件大小441MB。大部分为英文文本，少量中文文本。训练集：1255张 测试集：300

```
1  下载地址：http://www.cs-chan.com/source/ICDAR2017/totaltext.zip
2  文献：http:// arxiv.org/pdf/1710.10400v
```

Google FSNS(谷歌街景文本数据集)

该数据集是从谷歌法国街景图片上获得的一百多万张街道名字标志，每一张包含同一街道标志牌的不同视角，图像大小为600*150，训练集1044868张，验证集16150张，测试集20404张。

```
1  下载地址：http://rrc.cvc.uab.es/?ch=6&com=downloads
2  文献：http:// arxiv.org/pdf/1702.03970v1
```

COCO-TEXT

该数据集，包括63686幅图像，173589个文本实例，包括手写版和打印版，清晰版和非清晰版。文件

大小12.58GB，训练集：43686张，测试集：10000张，验证集：10000张

```
1  文献：http://arxiv.org/pdf/1601.07140v2
2  下载地址：https://vision.cornell.edu/se3/coco-text-2/
```

Synthetic Data for Text Localisation

在复杂背景下人工合成的自然场景文本数据。包含858750张图像，共7266866个单词实例，28971487个字符，文件大小为41GB。该合成算法，不需要人工标注就可知道文字的label信息和位置信息，可得到大量自然场景文本标注数据。

```
1  下载地址：http://www.robots.ox.ac.uk/~vgg/data/scenetext/
2  文献：http://www.robots.ox.ac.uk/~ankush/textloc.pdf
3  Code：https://github.com/ankush-me/SynthText (英文版)
4  Code https://github.com/wang-tf/Chinese_OCR_synthetic_data(中文版)
```

Synthetic Word Dataset

合成文本识别数据集，包含9百万张图像，涵盖了9万个英语单词。文件大小为10GB

```
1  下载地址：http://www.robots.ox.ac.uk/~vgg/data/text/
```

Caffe-ocr中文合成数据

数据利用中文语料库，通过字体、大小、灰度、模糊、透视、拉伸等变化随机生成，共360万张图片，图像分辨率为280x32，涵盖了汉字、标点、英文、数字共5990个字符。文件大小约为8.6GB

```
1  下载地址：https://pan.baidu.com/s/1dFda6R3
```

Reference

- 综述

  自然场景文本检测识别技术综述

  白翔:：图像OCR年度进展|VALSE2018之十一

白翔：趣谈“捕文捉字” — 场景文字检测 | VALSE2017之十

基于深度学习的目标检测及场景文字检测研究进展

知乎文本检测综述

优秀论文解读博客

知乎专栏:小石头的码疯窝

- 文本检测
  - CTPN

    场景文字检测—CTPN原理与实现

    CTPN: Tensorflow
  - EAST

    Bolg: EAST

    知乎：文本检测之EAST

    EAST：tensorflow

    EAST: Keras

    EAST: Advanced keras
  - SegLink

    SegLink_Blog

    文本检测之SegLink
  - PixelLink

    文本检测之PixelLink

Github: PixelLink

- TextBoxes

论文笔记：TextBoxes++: A Single-Shot Oriented Scene Text Detector

Github: TextBoxes++

- 角定位

基于角定位于区域分割

- 文本识别
  - ASTER

Github: ASTER

- TextSpotter
  - Mask TextSpotter

华科白翔教授团队ECCV2018 OCR论文：Mask TextSpotter

----------本文结束🐾知识分享，方便你我----------

深度学习 CV OCR

‹ AwsomeProcess 深度学习-TextDetection ›

昵称 邮箱 网址(http://)

Just go go

表情 | 预览

回复

1 评论

Anonymous Chrome 71.0.3578.98 Linux

2019-03-08 回复

ＣＴＰＮ不是华工金连文老师提出的吧

Powered By Valine

v1.3.9