**图像文本定位**

1. 目标
2. 实现图像文本行定位，以便输入基于文本行的ocr识别模型。
3. 扩展现有数据集，探索可批量生成带有背景的多语言混合图片的方式。
4. 数据处理与数据集扩展

2.1数据集处理

本实验数据集主要来自ICDAR2013数据集。数据集的初始格式是文本行四个顶点的坐标。为了符合本模型训练需求，将其细分成一个个宽度约为16像素的小块的n个顶点坐标。

2.2进行数据集扩展的背景

现有的文本检测数据集缺乏多语言混合的情况，无法在文本定位的同时对语言做分类。同时考虑到现在ocr定位模型普遍采用复杂背景下文字作为训练集，以满足各种场景下识别需求。因此希望寻找可批量生成带有背景的多语言混合图片的方式。

2.3获取背景图像中适合放置文本的位置信息

1.采取读入灰度图，若是浅色背景，将其转换为深色为背景的图片。

2．对图像进行膨胀操作，增强前景。

1. 有ostu算法进行二值化
2. 获取行连通区域，选取长度大于一定值得连通区域作为文本行。

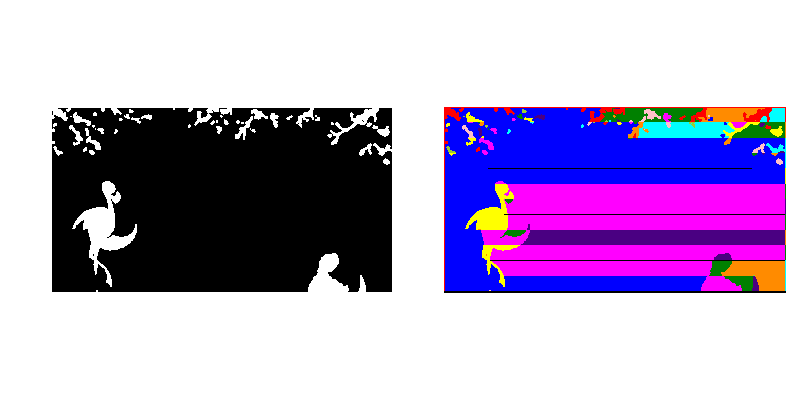


图1：文本行分割效果图

（同时尝试了canny边缘提取，获取凸包等方法，但效果均不佳，产生的连通区域过于细碎或不准确。）

2.4文本语料处理

中文语料从维基百科数据集获取。英文语料从美国当代英语语料库（COCA）获取。Latex公式则利用正则表达式从arXiv论文中提取。对latex公式进行标准化处理，如用标准表达式替换自定义函数，如

对于\newcommand{\uind}[2]{^{#1\_1 \, ... \, #1\_{#2}} }需要将‘^’替换为\uind，并将数字写成参数表形式。

2.5在图像上添加文字

根据2.3中获得的适合放置文字的文本行，可以很方便得在图像上添加中文和英文，但对于latex，需利用掩膜运算将latex公式与背景进行合成。并对添加文字方法进行优化，由于latex公式高度变化可能很大，需进行间距调整。同时跟据文字，背景图片信息及引入随机变量，如文字颜色，词句长短等。判断背景颜色深浅需要RGB转换为YUV，其中“Y”表示明亮度,y= R\*0.299 + G\*0.587 + B\*0.114.

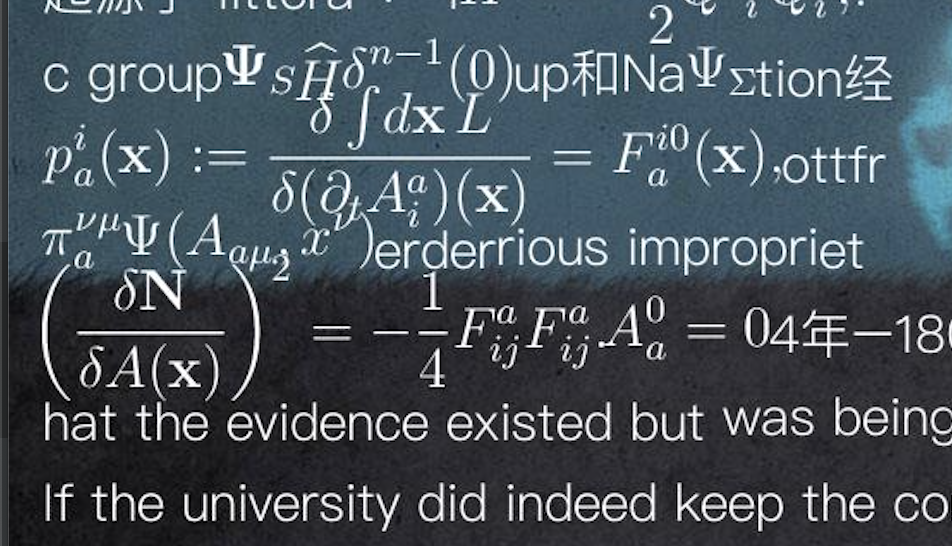


图2：优化前，latex公式行间距存在问题

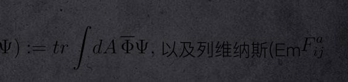


图3：优化前，文字与背景颜色存在问题

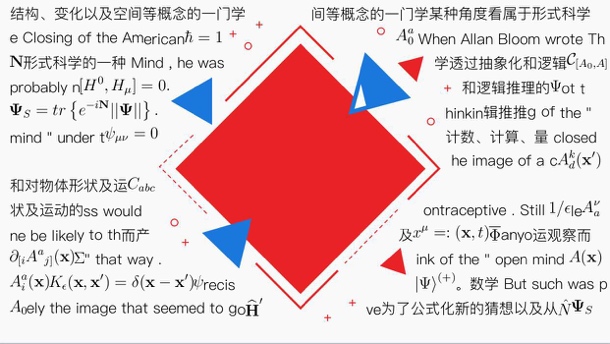
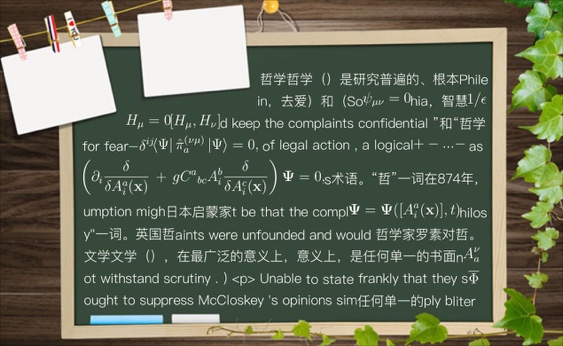
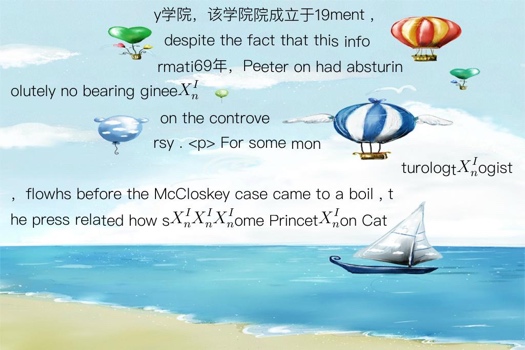
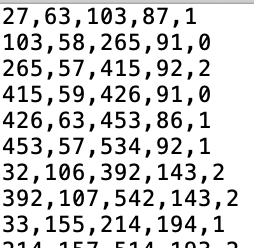
 

图4：生成图片与相应bbox坐标

三．文本检测模型

本实验采用基于faster rcnn的CTPN网络进行文本检测。文本检测可以看成特殊的目标检测，但文本检测要求较高，需要正确检出需要覆盖整个文本长度。具体实现流程包含三个部分检测小尺度文本框，循环连接文本框，文本行边细化。

3.1利用CNN与LSTM学习空间，序列特征

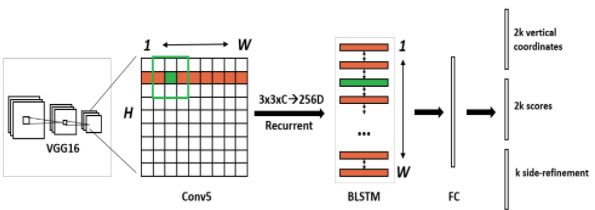


图5：用vgg16+lstm+fc学习空间，序列特征

-使用VGG16提取特征，得到conv5\_3的特征作为feature map，大小是W×H×C。

-在上述的feature map上使用大小为3\*s的滑动窗进行滑动，每个窗口都能得到一个长度为3×3×C的特征向量,这个特征向量将用来预测和9个anchor之间的偏移距离.

-将上一步得到的特征输入到一个双向的LSTM中，得到长度为W×256的输出.

-然后接入全连接层，得到N\*512\*W\*H的特征，准备输出。

3.2实现细节——垂直anchor

由于图片可能是不同尺寸和比例的，因此训练一个可以直接准确预测原始坐标的模型比较复杂。因此在图像上固定锚点，每个锚点对应一组anchor，这组anchor在之后目标位置的预测中用作参考anchor ，预测参考anchor的偏移量，和anchor score(是否包含文本)。这样只得到一些小数值的预测结果并挪动参考变量就可以达到更好的拟合结果。

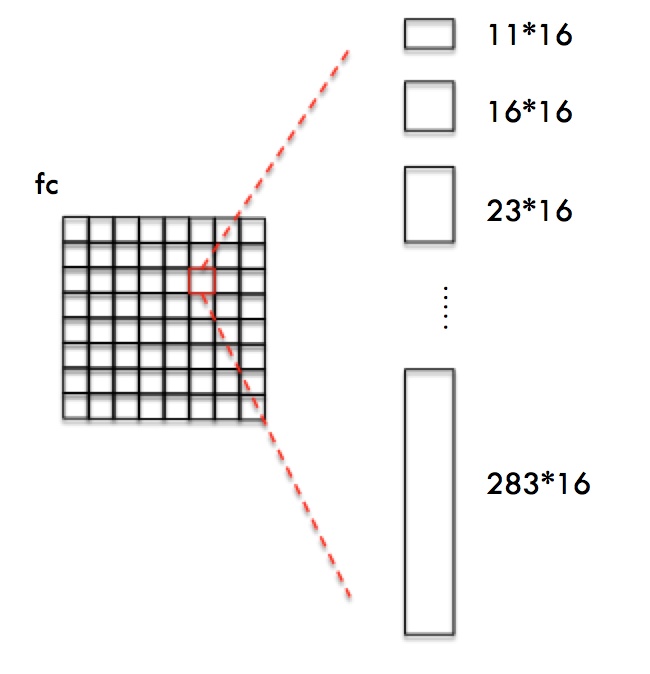


图7：垂直anchor

由于文本的长度是不固定，如果采用通用目标检测的方法，将会面临一个问题：如何生成好的text proposal．因此该模型只预测文本的竖直方向上的位置，水平方向的位置不预测。与faster rcnn中的anchor类似，但是不同的是，vertical anchor的宽度都是固定好的了，而高度从11像素到273像素(每次除以0.7)变化的。在卷积网络得到的feature map上一个值对应了原图像16个像素，每个点得到的特征向量反映的是该点与一组以这个点为中心的anchor之间的关系。

3.3区域建议网络(Region Proposal Networks)

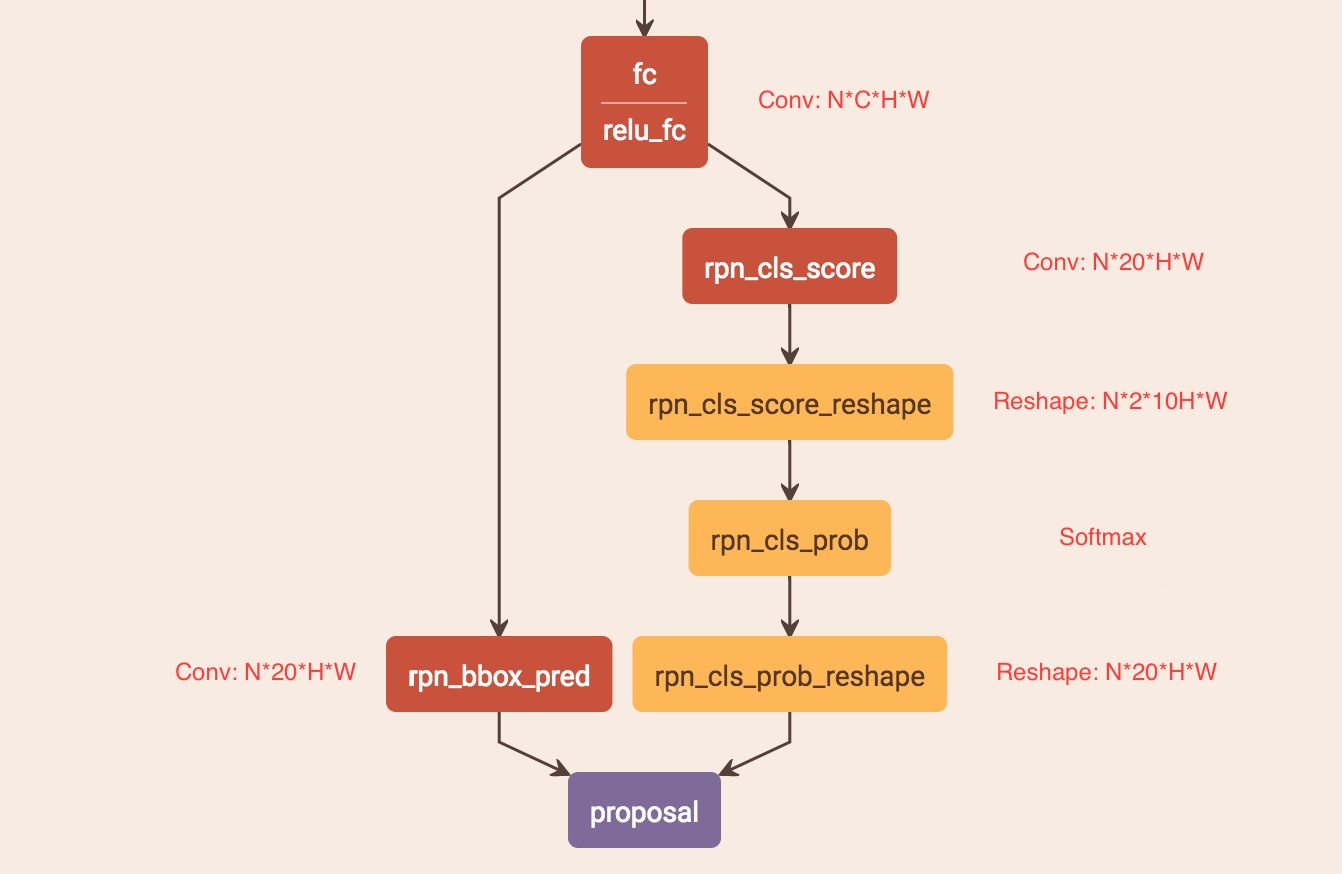


图7：区域建议网络

模型通过CNN和BLSTM学到一组“空间 + 序列”特征后，在"FC"后接入RPN网络，分为两个分支：

3.31 bounding box regression

由于 feature map每个点配备了9个Anchor，同时只回归中心y坐标与高度2个值。每个anchor 还有一个side-refinement值，这部分主要是用来精修文本行的两个端点的，表示的是每个proposal的水平平移量。所以bbox\_prediction分支有27个channels。

3.32 Softmax分类

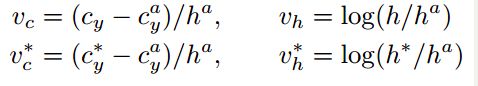
上图右边分支用于Softmax分类Anchor是否属于文本区。

* 1. Proposal Layer

Proposal Layer综合3个输入：fg/bg anchors分类器结果，对应的bbox regression的偏移量，side-refinement值. 计算出精准的proposal.

（1）Proposal Layer根据anchor的预测偏移量，side-refinement值,计算出预测anchor所在位置。

回归的高度和bounding box的中心的y坐标如下，带\*的表示是groundTruth，带a的表示是anchor。C是anchor中心，h是anchor高度。( Vc,Vh) 是回归预测的相对坐标， ( Vc\*,Vh\*)是Ground Truth。Cay 和 ha是Anchor的中心y坐标和高度，Cy和h是预测的中心坐标和高度。



当在两个水平侧的text proposals不能准确被一个ground truth文本行区域覆盖时，或者一些side proposals被丢弃时（例如：具有一个较低的text score），这会导致一个不精准的定位。这种不准确在通用目标检测中不是很严格，但在文本检测中不能忽视，尤其是对于那些小尺度文本行或词。为了解决该问题，引入side-refinement方法，可以准确地为每个anchor/proposal估计在水平左侧和水平右侧的offset。与y轴坐标的预测相似，计算了相对offset：

preview

Xside是相对于当前锚点最近水平侧（比如：左或右侧）的x预测坐标。Xside∗是ground truth侧在x轴坐标。Cxa是在x轴的锚点中心。Wa=16是锚点宽度。

（2）按照输入的foreground softmax scores由大到小排序anchors，提取前pre\_nms\_topN(12000)个anchors，即提取s修正位置后的foreground anchors。

(3)限定超出图像边界的foreground anchors为图像边界.

(4)剔除非常小的foreground anchors

(5)进行非最大值抑制（nonmaximum suppression），去除多余anchor。将所有框的得分排序，选中最高分及其对应的框。然后遍历其余的框，如果和当前最高分框的重叠面积(IOU)大于一定阈值，我们就将框删除。接着遍历其余的框，如果和当前最高分框的重叠面积(IOU)大于一定阈值，就将框删除。

（6）再次按照nms后的foreground softmax scores由大到小排序fg anchors，提取前post\_nms\_topN(2000)结果作为proposal输出。

## 3.4文本线构造

所有的text proposals只保留score大于0.7的proposal。每两个相近的proposal组成一个pair，合并不同的pair直到无法再合并为止（没有公共元素）。判断两个proposal，Bi和Bj组成pair的条件：

（1）Bj->Bi， 且Bi->Bj。（Bj->Bi表示Bj是Bi的最好邻居）

（2）Bj->Bi条件：Bj是Bi的邻居中距离Bi最近的，且该距离小于50个像素。Bj和Bi的vertical overlap大于0.7

## 3.5 模型输出和Loss functions

模型有三个outputs，它们会一起连接到最后的FC layer上：text/non-text scores(s)、垂直坐标(v=vc,vh)、side-refinement offset (o)。9个anchors在conv5中的每个空间位置上预测它们，在各自的output layer上产生2\*9, 2\*9和9个参数。

CTPN采用了多任务学习来联合优化模型参数。我们引入了三个loss functions：Lcls,Lrev,lreo，各自计算text/non-text score、坐标以及side-refinement的error。需要最小化总目标函数（L）：

L(si,vj,ok)=1Ns∑iLcls(si,s∗i)+λ1Nv∑jLrev(vj,v∗j)+λ2No∑kLreo(ok,o∗k)

其中，每个anchor是一个训练样本，i是minimatch中的一个anchor的索引。si是anchor i为一个真实文本的预测概率。Si\*=0,1是ground truth。j是一个关于y坐标回归中合法anchors集合的anchor索引。一个合法的anchor是一个已经定义的positive anchor（Sj∗=1），或者具有一个与ground truth的text proposal具有Intersection-over-Union(IoU) > 0.5 的重合度的anchor。vjvj和v∗jvj∗是第j个anchor相关的y坐标的prediction和ground truth。k是关于一个side-anchor的索引，side-anchor被定义成在距离ground truth文本行限定框左侧或右侧的水平距离内的一个anchors集合。ok和ok∗分别是在第k个anchor相关的在x轴上的predicted offsets 和ground truth offsets。Lrev和Lreo是regression loss。

四．训练效果与不足之处

经过30000step后，参数已基本收敛，在验证集上损失低于0.1。效果基本符合预期。但由于数据集扩展部分尚需完善，如在图像上添加噪音，对图像进行扭曲，以模仿拍摄图片，暂未进行语言分类工作。

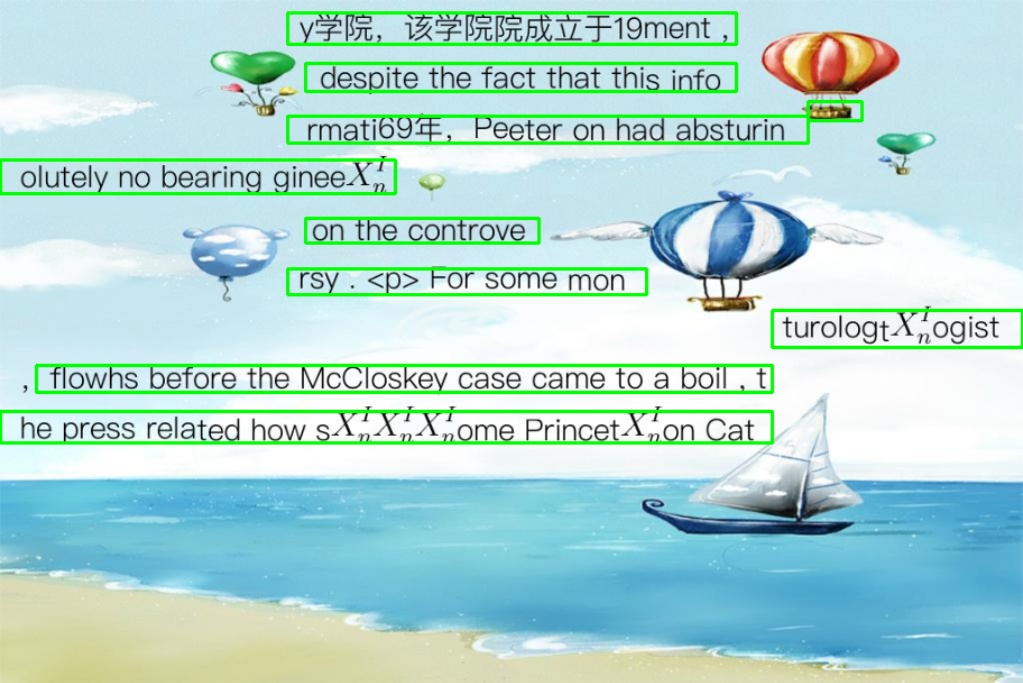
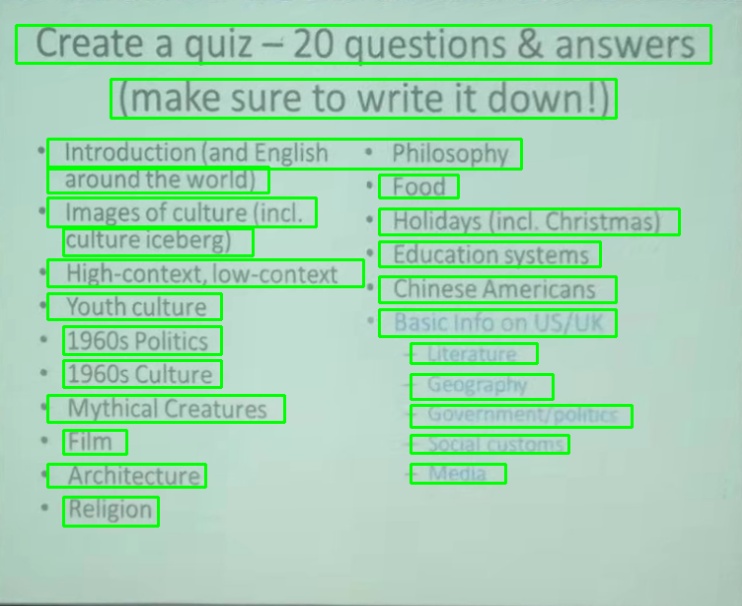
 

图8：左边为合成图像，右侧为拍摄图像

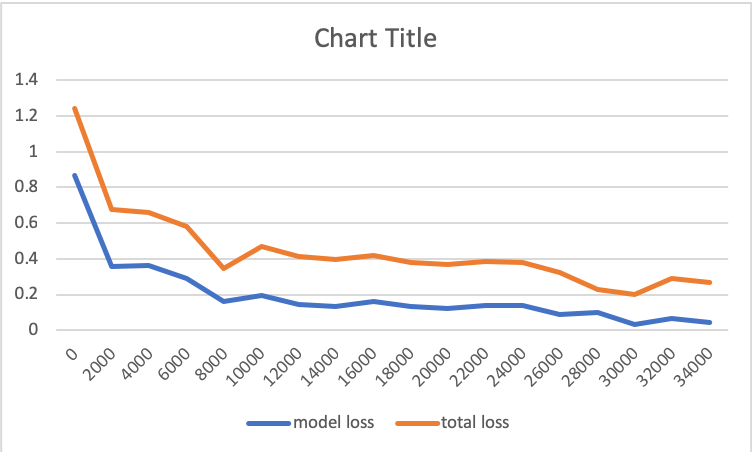


图9：模型损失