Detecting Oriented Text in Natural Images by Linking Segments

论文阅读

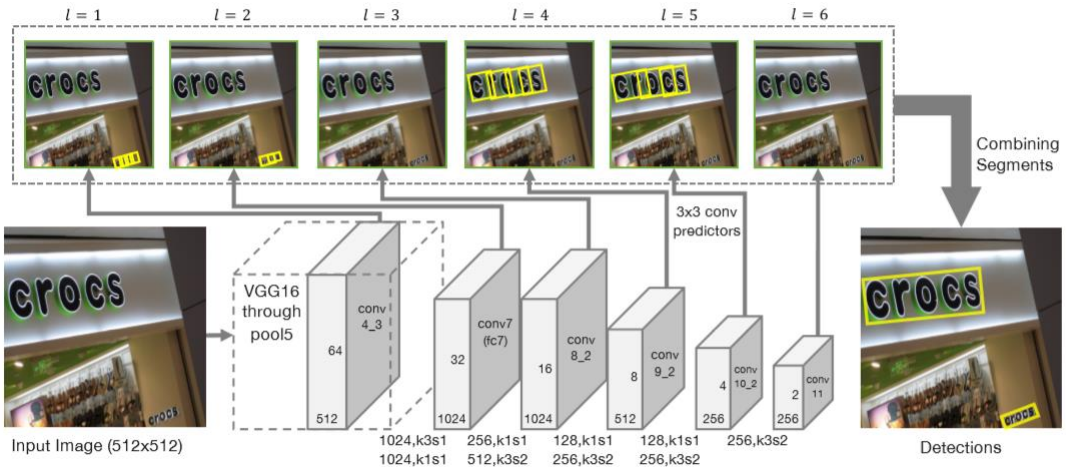
网络大致结构

网络结构就是 SSD 的结构都是在不同的 6 个特征层进行预测操作

网络的 input 和 output ：

首先是输入，因为网络全部采用卷积结构，所有对输入图片大小没有要求，可以是任意大小和任意长宽比，这点比较好理解。这里假定输入图片大小为 $W_I \times H_I$ 。

然后是输出，输出为 segments 和 links。segments 可以理解为是一个一个的小框，这些小框类似于 SSD 中的 default boxes，它们不一定一个框能框一个字，可能就框一个字的一部分。一个 segment 用公式 $b=(x_b,y_b,w_b,h_b,\theta_b)$表示，其中 x_b,y_b 表示 segment 的中心，w_b,h_b 表示 segment 的宽和高，θ_b 表示该 segment 的旋转角。links 就是将 segments 连接起来，意思就是就是两个框是不是同一个文本的一个概率值。



Segment 的检测方法

因为提取出 6 层 feature map，每层都需要输出 segments，segments 的表示方法为 $b=(xb,yb,wb,hb,\theta b)$。检测一个 segment 那么网络需要输出 segment 的置信度和 segment 相对于 default boxes 的五个回归偏移量。文章中是 7 个 channel，前两个通道是 soft-max 之后 segment 的置信度，后五个是偏移量和角度。

对于第 L 层的 feature map(W_l,H_l)来说，一个点在 feature map 上的坐标为(x,y)，对应原图是坐标为(x_a,y_a)的点，那一个 default box 的中心坐标为(x_a,y_a)，由下面的式子表示：

$$x_a = \frac{w_I}{w_l}(x+0.5); \quad y_a = \frac{h_I}{h_l}(y+0.5)$$

网络会预测出 segment 相对于对应位置的 default box 的五个偏移量 $(\Delta x_s, \Delta y_s, \Delta w_s, \Delta h_s, \Delta \theta_s)$，和上面 default box 的中心坐标($x_a$,y_a)

$x_s=a_l\Delta x_s+x_a$

$y_s=a_l\Delta y_s+y_a$

$w_s=a_l exp(\Delta w_s)$

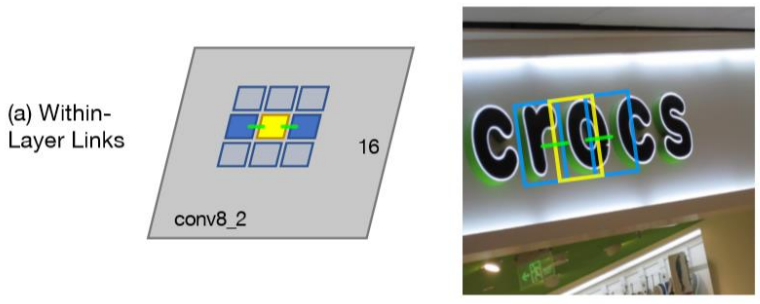
$h_s=a_l exp(\Delta h_s)$

$\theta_s=\Delta\theta_s$

$a_l = \gamma \frac{w_I}{w_l}$,这里$\gamma = 1.5$

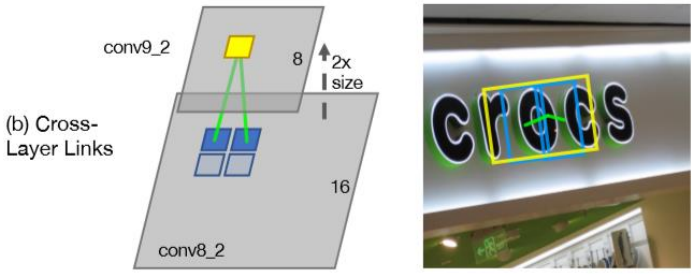
link 的检测方法

（1）层内的 Link 的检测
一个 link 连接着相邻两个 segment，表示他们是属于同一个字或者在同一框中。link 的作用不仅是将相邻的 segment 连接起来，还可以区分邻近的 segment 但是不属于同行或者同一个标定框。
检测 link 使用的 feature map 与检测 segment 使用的是同一 feature map，所以对于同一层 feature map 来说，一个 segment 有其他 8 个相邻的 segment，那 links 就是每个 feature map 经过卷积后输出 16 个通道，每两个通道表示 segment 与邻近的一个 segment 的 link。



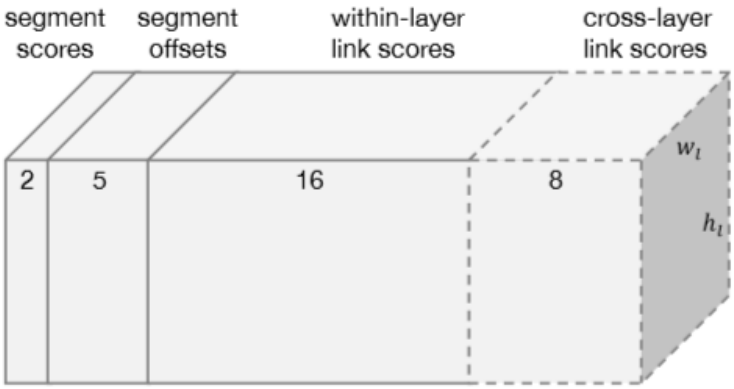
（2）跨层的 Link 检测
Cross-Layer Link 连接的是相邻两层 feature map 产生的 segments，比如，1-th 层（即 conv4_3）的 feature map 和 2-th 层（即 conv7）的 feature map 产生的 segments 通过 Cross-Layer Link 连接。
这个网络有个重要的属性方便我们进行 Cross-Layer Link 连接，就是我们提取出来的 6 个 feature map 中，上一层的大小是下一层的四倍（长宽各两倍）。但是值得注意的是，只有 feature map 是偶数的时候才满足这个属性，所以在实际操作中，输入图像的长宽大小都要缩放到 128 的整数倍。例如，一张 1000×800 的图片，首先会先缩放到 1024×768 大小。

由于上层的 feature map 为下一层的四倍，那相当于一个 segment 与另一层的四个 segment 相邻。这时除 1-th 的 feature map 外，其他五个 feature map 每个经过卷积后都要输出 8 个通道，每两个通道表示一个 Cross-Layer Link。



那么每层 feature map 的输出

经过上述介绍，每层 feature map 提取出来后，还要经过卷积出来，最后输出的有 segments 的信息也有 links 的信息，一共有 31 个 channel



但是第一层也就是 con4_3 层 f，eature map 的输出少 Cross-Layer Link。

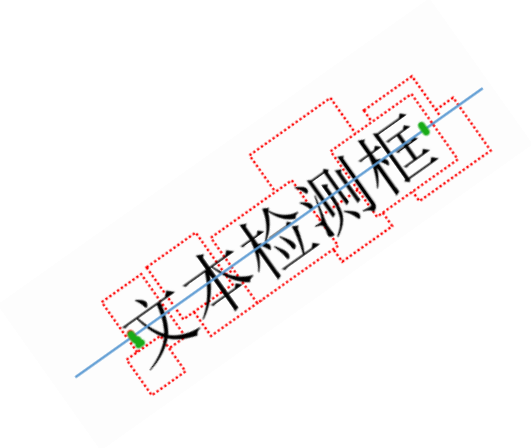
利用 links 将 segments 连接

将 segments 看作结点，links 看作边，下面算法的输入是 segments 和 links 构成的一张图，每个文本框看作一个连通分量。

Algorithm 1 Combining Segments

Input: $\mathcal{B} = \{s^{(i)}\}_{i=1}^{|\mathcal{B}|}$ is a set of segments connected by links, where $s^{(i)} = (x_s^{(i)}, y_s^{(i)}, w_s^{(i)}, h_s^{(i)}, \theta_s^{(i)})$.
Find the average angle $\theta_b := \frac{1}{|\mathcal{B}|}\sum_{\mathcal{B}} \theta_s^{(i)}$.
For a straight line $(\tan\theta_b)x + b$, find the b that minimizes the sum of distances to all segment centers $(x_s^{(i)}, y_s^{(i)})$.
Find the perpendicular projections of all segment centers onto the straight line.
From the projected points, find the two with the longest distance. Denote them by (x_p, y_p) and (x_q, y_q).
$x_b := \frac{1}{2}(x_p + x_q)$
$y_b := \frac{1}{2}(y_p + y_q)$
$w_b := \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} + \frac{1}{2}(w_p + w_q)$
$h_b := \frac{1}{|\mathcal{B}|}\sum_{\mathcal{B}} h_s^{(i)}$
$b := (x_b, y_b, w_b, h_b, \theta_b)$
Output: b is the combined bounding box.

1. 将连接后的所有结果作为输入，将连接在一起的 segments 当作是一个小的集合，称为 B。
2. 将 B 集合中所有 segment 的旋转角求平均值作为文本框的旋转角称为 θ_b
3. 将旋转角求 $\tan\theta_b$ 作为斜率，这样就可以得到一系列的平行线，求得 B 集合中所有 segment 的中心点到直线距离的和最小的那条直线。
4. 将 B 集合中所有 segment 的中心点垂直投影到 3 步骤中找到的直线上。
5. 在投影中找到距离最远的两个点称为(xp,yp)和(xq,yq)
6. 上述两点的均值作为框的中心点，宽为上述两点的距离，高为 B 集合中所有 segment 的高的均值。



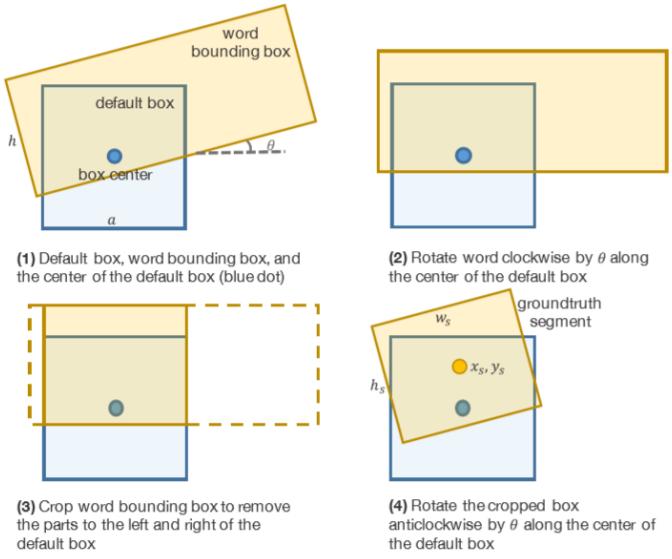
segments 和 links 标签的生成

在求 segments 和 links 的标签前先确定与其对应的 default box 的标签值。

1) the center of the box is inside the word bounding box;
2) the ratio between the box size a_l and the word height h satisfies:

$$\max(\frac{a_l}{h}, \frac{h}{a_l}) \leq 1.5 \qquad (9)$$

原生数据只是给了四边形的四个点。
只要 default box 的中心点在标定的文本框内，还有两个框高的比例满足上述关系，就认为这个 default box 为正样本。
接下来计算位置偏移量和角度的标签值



1.选择一个正样本的 default box，如图中蓝框所示，其中的蓝点是 default box 的中心点

2.将文本框顺时针旋转为 θ，使其成为水平框
3.在得到的水平框能截取 default box 大小的区域（长为 default box 的长，高仍为文本框的高）
4.根据截取后的水平框，沿着其中心点逆时针旋转 θ 角

这时候 w_s,h_s,x_s,y_s,θ_s 就知道了

$$x_s=a_l\Delta x_s+x_a$$

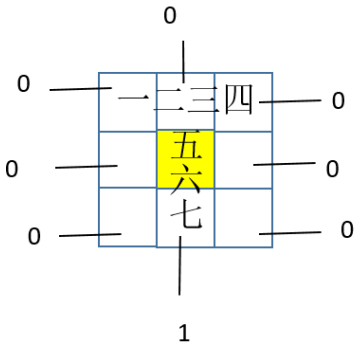
$$y_s=a_l\Delta y_s+y_a$$

$$w_s=a_l exp(\Delta w_s)$$

$$h_s=a_l exp(\Delta h_s)$$

$$\theta_s=\Delta\theta_s$$

link（包括 within-layer link 和 cross-layer link）的标签值，满足下面两个条件：
1. link 连接的两个 default box 都为正样本
2. 两个 default box 属于同一个文本框



损失函数的定义

损失函数定义如下式所示：

$$L(\mathbf{y}_s,\mathbf{c}_s,\mathbf{y}_l,\mathbf{c}_l,\hat{\mathbf{s}},\mathbf{s})=\frac{1}{N_s}L_{\text{conf}}(\mathbf{y}_s,\mathbf{c}_s)+\lambda_1\frac{1}{N_s}L_{\text{loc}}(\hat{\mathbf{s}},\mathbf{s})+\lambda_2\frac{1}{N_l}L_{\text{conf}}(\mathbf{y}_l,\mathbf{c}_l)$$

$\mathbf{y}_s,\mathbf{c}_s$ 分别表示 segment 的标签值和预测值，$\hat{\mathbf{s}},\mathbf{s}$ 分别表示偏移量的预测值和标签值，$\mathbf{y}_l,\mathbf{c}_l$ 分别表示 link 的预测值和标签值。N_s 为图像中所有正样本的 default boxes 的个数。N_L 为图像中所有正样本的 links 的个数。λ_1 和 λ_2 作者都设为 1。