论文题目

Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation

提出前提

水平和多方向的 text 已经能被很好的 detect 出来,但是弯曲文本目前来说还是很有挑战性

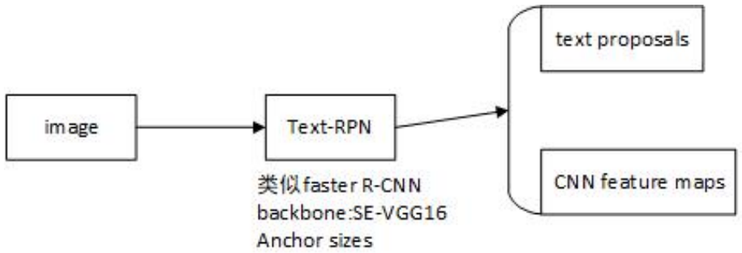
本文的观点

提出了一种自适应文本区域表示的文本检测方法,简单点说就是表征文本框的 points 个数是不定的,不同的文本 points 的个数不同

本文的做法(two stage)

用 Text-RPN 提取候选框之后,每一个文本区域用 RNN 验证改善来预测自适应数目的边界点

1. 输入图片通过 Text-RPN 得到 text proposals+CNN feature maps



Layer	Kernel
Conv1	$[3 \times 3, 64] \times 2$
Pool1	2×2, stride 2
SE1	4, 64
Conv2	$[3 \times 3, 128] \times 2$
Pool2	2×2, stride 2
SE2	8, 128
Conv3	$[3 \times 3, 256] \times 3$
Pool3	2×2, stride 2
SE3	16, 256
Conv4	$[3 \times 3, 512] \times 3$
Pool4	2×2, stride 2
SE4	32, 512
Conv5	$[3 \times 3, 512] \times 3$
Pool5	2×2, stride 2
SE5	32, 512

Table 1. The architecture of SE-VGG16 network. For SE block, its kernel means the channel numbers of the two FC layers in it

SE block 对效果的提升

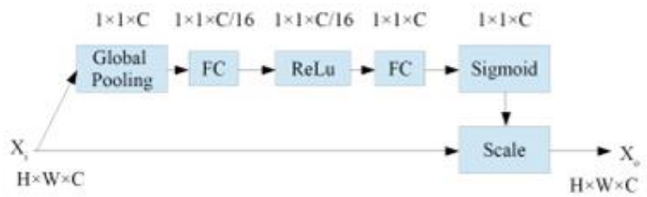


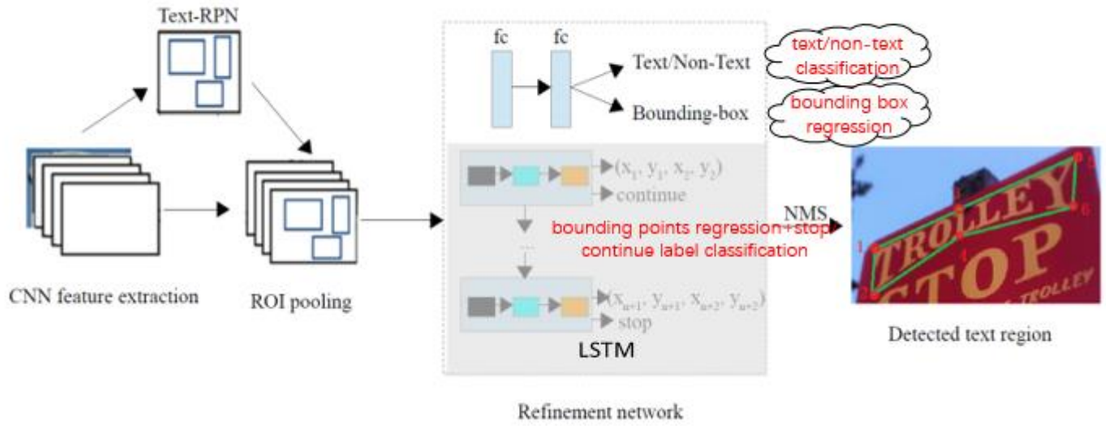
Figure 4. The architecture of SE block.

SE block能产生显著的性能改进

Backbone	Recall	Precision	Hmean
CTW1500			
VGG16	79.1	79.7	79.4
SE-VGG16	80.2	80.1	80.1
ICDAR2015			
VGG16	83.3	90.4	86.8
SE-VGG16	86.0	89.2	87.6

Table 2. Ablation study on backbone network.

2. text proposals+CNN feature maps 通过 proposal refinement(RNN:LSTM)得到 text/non-text classification+bounding box regression+bounding points regression+stop/continue label classification
最后在通过 NMS(适用于多边形的 NMS)得到最终的 detect points
LSTM 的工作是:bounding points regression+stop/continue label classification,最终结果导致不同的文本 points 的数目不同(adaptive)



最终效果

Method	Recall	Precision	Hmean
SegLink [24]	40.0	42.3	40.8
EAST [34]	49.1	78.7	60.4
DMPNet [16]	56.0	69.9	62.2
CTD [17]	65.2	74.3	69.5
CTD+TLOC [17]	69.8	77.4	73.4
TextSnake [19]	**85.3**	67.9	75.6
Proposed	80.2	**80.1**	**80.1**

Table 4. Results on CTW1500.

Method	Recall	Precision	Hmean
EAST [34]	67.4	87.3	76.1
SegLink [24]	70.0	86.0	77.0
PixelLink [3]	73.2	83.0	77.8
TextSnake [19]	73.9	83.2	78.3
InceptText [28]	79.0	87.5	83.0
MCN [18]	79.0	**88.0**	83.0
Proposed	**82.1**	85.2	**83.6**

Table 8. Results on MSRA-TD500.

Method	Recall	Precision	Hmean
SegLink [24]	23.8	30.3	26.7
EAST [34]	36.2	50.0	42.0
DeconvNet [2]	44.0	33.0	36.0
Mask Textspotter [20]	55.0	69.0	61.3
TextSnake [19]	74.5	**82.7**	78.4
Proposed	**76.2**	80.9	**78.5**

Table 5. Results on TotalText.

Method	Recall	Precision	Hmean
TextBoxes [12]	83.0	88.0	85.0
SegLink [24]	83.0	87.7	85.3
He *et al.* [5]	81.0	92.0	86.0
Lyu *et al.* [21]	84.4	92.0	88.0
FOTS [15]	-	-	88.2
RRPN [22]	87.9	94.9	91.3
FEN [31]	89.1	93.6	91.3
Mask Textspotter [20]	88.6	**95.0**	91.7
Proposed	**89.7**	93.7	**91.7**

Table 6. Results on ICDAR2013.

Method	Recall	Precision	Hmean
SegLink [24]	76.8	73.1	75.0
RRPN [22]	77.0	84.0	80.0
He *et al.* [5]	81.0	92.0	86.0
R2CNN [8]	79.7	85.6	82.5
TextSnake [19]	80.4	84.9	82.6
PixelLink [3]	82.0	85.5	83.7
InceptText [28]	80.6	90.5	85.3
Mask Textspotter [20]	81.0	**91.6**	86.0
Proposed	**86.0**	89.2	87.6
FOTS [15]	-	-	**88.0**

Table 7. Results on ICDAR2015.

相对于其他做任意文本检测的算法速度要更快

Method	Scale	Speed
TextSnake [19]	768	1.1 fps
Mask Textspotter [20]	720	6.9 fps
Proposed	720	**10.0 fps**

Table 9. Speed compared on different detection methods supporting arbitrary shape texts.

改善方向

1. 任意形状的文本可以用角点来检测,对训练图像来说更容易标注
2. 可以加上文本识别的,需要考虑端到端的场景文本检测识别