Text Detection with Selected Anchors

Anna Zhu , Hang du , Shengwu Xiong
School of Computer Science and Technology, Wuhan University of Technology, China

Introduction

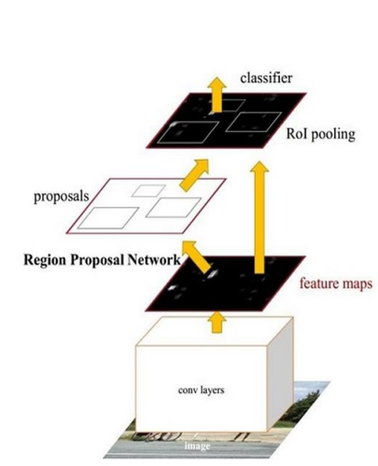
Text information in natural scenes provides important clue for content-based image analysis. It's the initial and promising step for its accuracy has extremely influence on the sequential text recognition.

Contributions:

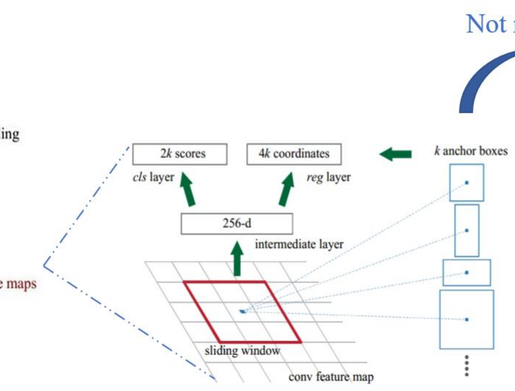
(a) Proposing a novel proposal extraction scheme with learnable anchors instead of predefined ones.
(b) The anchor selection region proposal network (AS-RPN) has the ability to predict non-uniform and arbitrary shape with significant reduced anchors.
(c) The scheme can detect text instances with arbitrary orientation.

Motivation

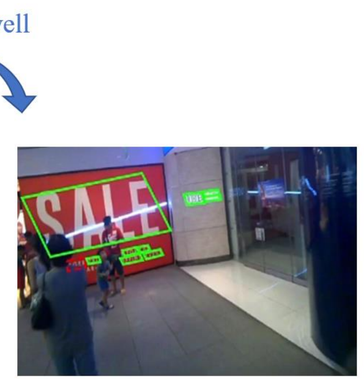
Predefined anchor scheme are not efficient and accurate enough for different text instances detection.



Architecture of Faster RCNN



Predefined dense anchors

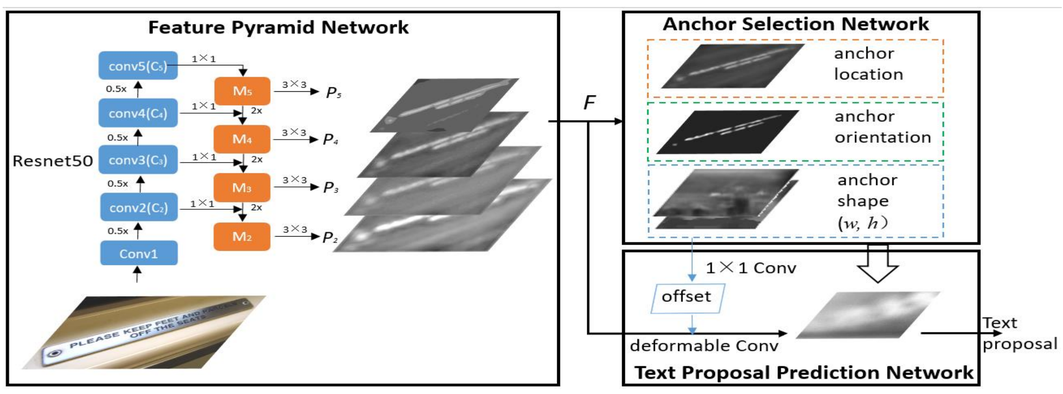


Instance with different scales and orientations

Solve the problem of arbitrary shape and orientation by using learnable anchors instead of fixing them.

Proposed Method

The proposed network



Architecture

The architecture is composed of three components, namely the Feature Pyramid Network (FPN), anchor selection network and text prediction network.

The inputs of the network are the whole images and directly outputs regressed text bounding boxes and confidences from the convolutional features, referring to a set of predicted anchors with arbitrary locations, orientations and shapes of width and height.

Training

The loss function can be divided into three parts:

$$L = L_{conf} + L_{reg} + L_{anchor}$$

The anchor loss can be further divided into three parts:

$$L_{anchor} = \alpha L_{loc} + \beta L_{angle} + \lambda L_{shape}$$

$$L_{loc} = \begin{cases} -\alpha(1-y')^{\gamma}\log y', y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), y = 0 \end{cases}$$

$$L_{angle} = 1 - \cos(\hat{\theta} - \theta_g)$$

$$L_{shape} = L_1\left(1 - min\left(\frac{w}{w_g}, \frac{w_g}{w}\right)\right) + L_1\left(1 - min\left(\frac{h}{h_g}, \frac{hg}{h}\right)\right)$$

Note: α, β, λ are parameters to balance the location, orientation and shape prediction branches which are set to $\alpha = \beta = 1$; $\lambda = 0.1$; $*_g$ donates the *(angle and shape) target.

Prediction

The location branch outputs a probability map, which indicates the probability of the text center existing at that location.

The orientation prediction branch outputs a soft map assigned with to a value within the range of [0,1].

The shape prediction branch is performed on feature map of each level and predicts the best shape (w, h) which has the highest IoU with the ground truth bounding box at each location estimated by:

$$w = k \times s \times exp(dw), h = k \times s \times exp(dh)$$

Experimental Results

Dataset: COCO-Text, ICDAR 2013, ICDAR 2015, MSRA-TD 500

The number of the proposals with different anchor-based methods

Example proposals of RPN (top row) and AS-RPN (bottom row)

Measure \ Method		RPN	FPN-RPN	AF-RPN	AS-RPN
IoU_0.5	TR_{50}	67.2	67.5	73.3	**74.5**
	TR_{100}	76.9	77.2	81.8	**82.9**
	TR_{300}	86.6	87.4	**89.3**	88.6
IoU_0.75	TR_{50}	22.8	28.8	35.0	**36.2**
	TR_{100}	27.9	36.0	41.3	**44.6**
	TR_{300}	33.8	47.2	48.2	**48.8**
IoU_Avg	TR_{50}	30.6	33.5	38.2	**38.8**
	TR_{100}	35.9	39.8	43.6	**44.9**
	TR_{300}	41.7	48.0	49.2	**50.0**

Region proposal quality evaluation on COCO-Text validation set

Comparison results with relevant approaches on ICDAR and MSRA-TD500

Approach	P (%)	R (%)	F (%)
CPTN[35]	74.22	51.56	60.85
Seg Link[27]	74.74	76.50	75.61
SSTD[41]	80.23	73.86	76.91
RRPN[26]	82.02	73.00	77.05
EAST*[38]	84.36	81.27	82.79
R2CNN[42]	85.62	79.68	82.54
Text boxes++[25]	87.80	78.50	82.90
Ours	**83.34**	**79.99**	**81.63**

Results on ICDAR 2015

Approach	P (%)	R (%)	F (%)
Baseline	57.40	54.50	55.90
He et al[37]	76.40	61.42	68.76
EAST*[38]	81.23	63.27	75.54
RRPN[26]	68.00	82.00	74.00
TextSnake[39]	83.20	73.90	78.30
Pixel Link[17]	83.00	73.20	77.82
Lyu et al[28]	87.60	76.20	81.50
Ours	**84.67**	**80.37**	**82.49**

Results on MSRA-TD 500

Detect results of selecting anchors for different tasks

Different shapes and scales of text instances

Arbitary orientations of text instances

Conclusion

This paper present an accurate scene text detection approach named AS-RPN which can generate high-quality text proposals through anchor location prediction, anchor orientation estimation and anchor shape prediction branches.