FOTS: Fast Oriented Text Spotting with a Unified Network

Xuebo Liu[1], Ding Liang[1], Shi Yan[1], Dagui Chen[1], Yu Qiao[2], and Junjie Yan[1]
[1]SenseTime Group Ltd.
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

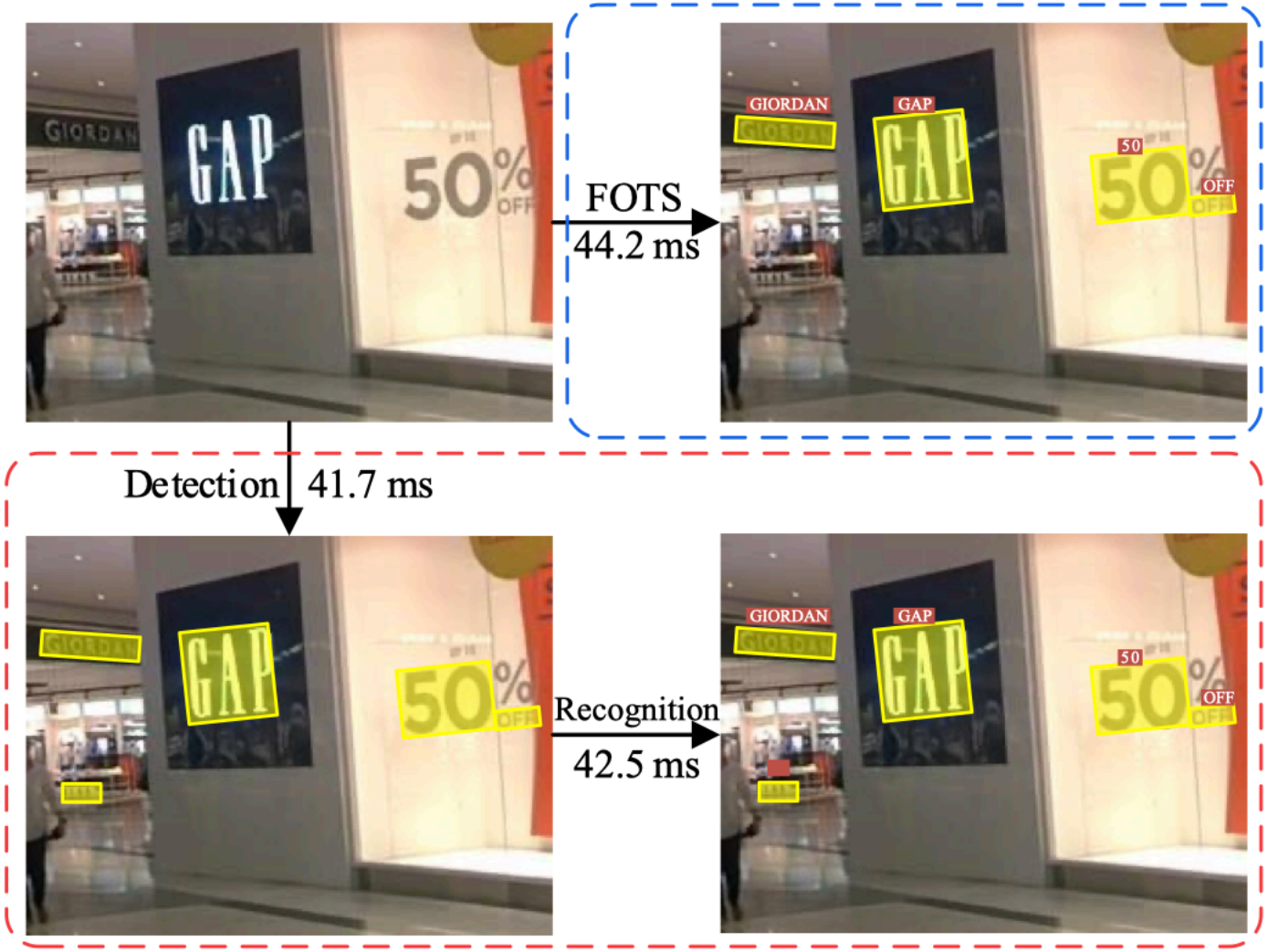
https://arxiv.org/abs/1801.01671
Accepted by CVPR 2018

- Introduction
- Overall architecture
- Loss function
- Conclusion

Introduction

1.速度比较，相较于检测+识别的两个单独的任务，把它们联合在一起使用速度更快

2.另一个问题是两个任务分开完成忽略了在检测和识别中共享的视觉线索中的相关性。单个检测网络不能由来自文本识别的标签监督。



Overall architecture

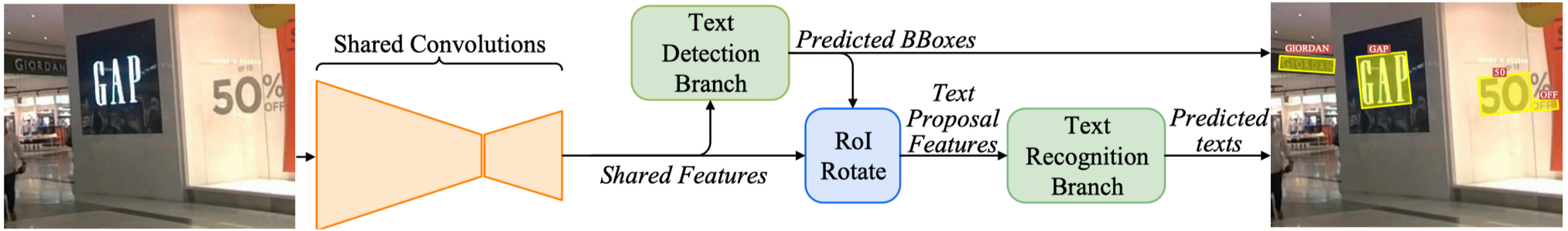


Figure 2: Overall architecture. The network predicts both text regions and text labels in a single forward pass.

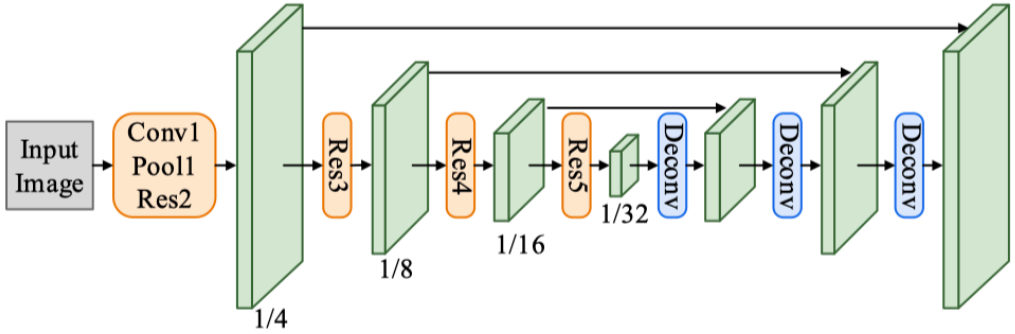


Figure 3: Architecture of shared convolutions. Conv1-Res5 are operations from ResNet-50, and Deconv consists of one convolution to reduce feature channels and one bilinear upsampling operation.

Type	Kernel [size, stride]	Out Channels
conv_bn_relu	[3, 1]	64
conv_bn_relu	[3, 1]	64
height-max-pool	[(2, 1), (2, 1)]	64
conv_bn_relu	[3, 1]	128
conv_bn_relu	[3, 1]	128
height-max-pool	[(2, 1), (2, 1)]	128
conv_bn_relu	[3, 1]	256
conv_bn_relu	[3, 1]	256
height-max-pool	[(2, 1), (2, 1)]	256
bi-directional_lstm		256
fully-connected		$\|S\|$

text detection branch

- 使用一层卷积进行输出6个通道，1个通道是score map，计算每个像素表示对应于原图中像素为文字的概率值。4个通道预测它到包含这个像素点的边界框的上、下、左、右边的距离。1个通道预测边界框的方向。正样本应用阈值过滤和NMS，可以得到最后的检测结果。

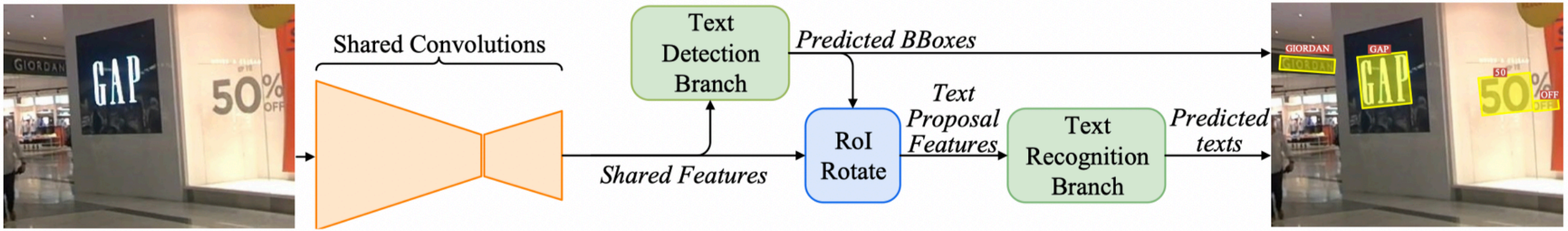


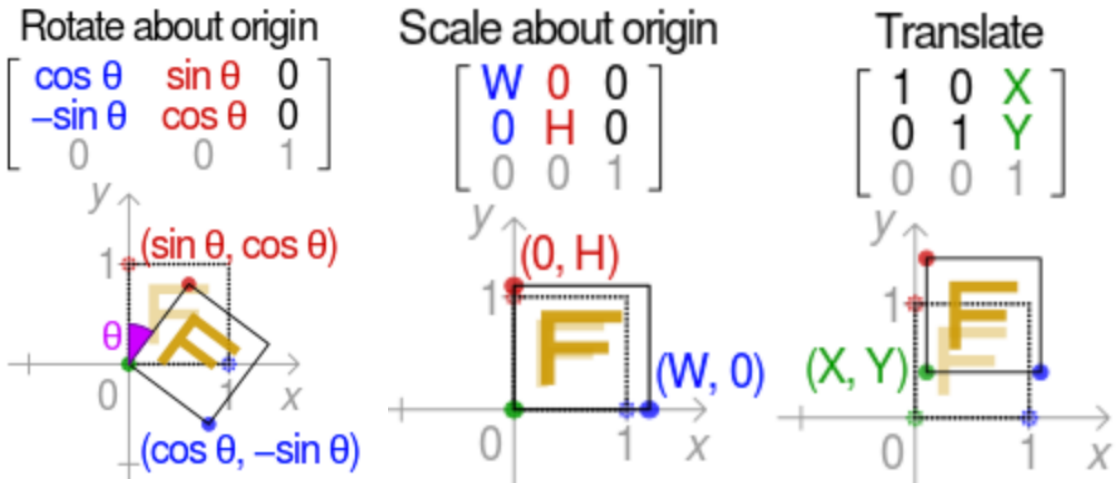
Figure 2: Overall architecture. The network predicts both text regions and text labels in a single forward pass.

RoIRotate

RoIRotate将原本多方向的边界框进行仿射变换到x，y标准坐标系中。这样一来端到端的训练就变成了可能。对每个ground-true区域分别使用仿射变换到共享特征图上，获得文本区域标准的水平特征图。

$$\mathbf{M} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

$$= s \begin{bmatrix} \cos\theta & -\sin\theta & t_x\cos\theta - t_y\sin\theta \\ \sin\theta & \cos\theta & t_x\sin\theta + t_y\cos\theta \\ 0 & 0 & \frac{1}{s} \end{bmatrix} \quad (8)$$



Loss function

- Loss$_{\text{detection}}$

$$L_{\text{cls}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{H}(p_x, p_x^*)$$

$$= \frac{1}{|\Omega|} \sum_{x \in \Omega} (-p_x^* \log p_x - (1 - p_x^*) \log(1 - p_x))$$

$$L_{\text{reg}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{IoU}(\mathbf{R}_x, \mathbf{R}_x^*) + \lambda_\theta (1 - \cos(\theta_x, \theta_x^*))$$

- Loss $_{\text{recognition}}$

$$L = L_{\text{detect}} + \lambda_{\text{recog}} L_{\text{recog}}$$

Training

- Data augmentation

- OHEM

Data augmentation is important for robustness of deep neural networks, especially when the number of real data is limited, as in our case. First, longer sides of images are resized from 640 pixels to 2560 pixels. Next, images are rotated in range $[-10°, 10°]$ randomly. Then, the heights of images are rescaled with ratio from 0.8 to 1.2 while their widths keep unchanged. Finally, 640×640 random samples are cropped from the transformed images.

As described in Sec. 3.2, we adopt OHEM for better performance. For each image, 512 hard negative samples, 512 random negative samples and all positive samples are selected for classification. As a result, positive-to-negative ratio is increased from 1:60 to 1:3. And for bounding box regression, we select 128 hard positive samples and 128 random positive samples from each image for training.

Conclusion

Contribution:
1.共享卷积特征
2.引入ROIRotate
3.速度和精度都优于之前的方法

Method	Detection			Method	End-to-End			Word Spotting		
	P	R	F		S	W	G	S	W	G
SegLink [43]	74.74	76.50	75.61	Baseline OpenCV3.0+Tesseract [26]	13.84	12.01	8.01	14.65	12.63	8.43
SSTD [13]	80.23	73.86	76.91	Deep2Text-MO [51, 50, 20]	16.77	16.77	16.77	17.58	17.58	17.58
WordSup [17]	79.33	77.03	78.16	Beam search CUNI+S [26]	22.14	19.80	17.46	23.37	21.07	18.38
RRPN [39]	83.52	77.13	80.20	NJU Text (Version3) [26]	32.63	-	-	34.10	-	-
EAST [53]	83.27	78.33	80.72	StradVision_v1 [26]	33.21	-	-	34.65	-	-
NLPR-CASIA [15]	82	80	81	Stradvision-2 [26]	43.70	-	-	45.87	-	-
R^2CNN [25]	85.62	79.68	82.54	TextProposals+DictNet [7, 19]	53.30	49.61	47.18	56.00	52.26	49.73
CCFLAB_FTSN [4]	88.65	80.07	84.14	HUST_MCLAB [43, 44]	67.86	-	-	70.57	-	-
Our Detection	88.84	82.04	85.31	Our Two-Stage	77.11	74.54	58.36	80.38	77.66	58.19
FOTS	91.0	85.17	87.99	FOTS	81.09	75.90	60.80	84.68	79.32	63.29
FOTS RT	85.95	79.83	82.78	FOTS RT	73.45	66.31	51.40	76.74	69.23	53.50
FOTS MS	**91.85**	**87.92**	**89.84**	FOTS MS	**83.55**	**79.11**	**65.33**	**87.01**	**82.39**	**67.97**

FOTS RT : resnet 50 -> resnet 34
FOTS MS: Multi-scale training

Dataset	Method	Speed		Params
		Detection	End-to-End	
IC15	Our Two-Stage	7.8 fps	3.7 fps	63.90 M
	FOTS	7.8 fps	7.5 fps	34.98 M
	FOTS RT	24.0 fps	22.6 fps	28.79 M
IC13	Our Two-Stage	23.9 fps	11.2 fps	63.90 M
	FOTS	23.9 fps	22.0 fps	34.98 M

Test on TITAN-XP GPU

FOTS: Fast Oriented Text Spotting with a Unified Network

CVPR 2018 Submission #1699

Thank you !