Focusing Attention: Towards Accurate Text Recognition in Natural Images

Zhanzhan Cheng, Gang Zheng and Shiliang Pu
Hikvision Research Institute
{chengzhanzhan, zhenggang3, pushiliang}@hikvision.com

Fan Bai and Shuigeng Zhou
Fudan University
{baif13, sgzhou}@fudan.edu.cn

Yunlu Xu
Shanghai Jiaotong University
xuyunlu1030@163.com

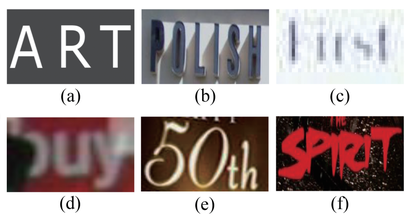
Abstract

- Scene text recognition has been a hot research topic in computer vision due to its various applications.
- Attention-based methods cannot get accurate alignments between feature areas and targets for complicated images; We call this phenomenon "attention drift".
- We propose the FAN (the abbreviation of Focusing Attention Network) method that employs a focusing attention mechanism to automatically draw back the drifted attention.
- Different from the existing methods, we adopt a ResNet-based network to enrich the deep representations of scene text images.
- Extensive experiments on various benchmarks, including the IIIT5k, SVT and ICDAR datasets, show that the proposed FAN method substantially outperforms the existing methods.

Motivation

Problem Statement

- In real scene text recognition tasks, many images (right picture) are complicated (e.g. distorted or overlapping characters, characters of different fonts, sizes and colors, and complex backgrounds) or low quality (illumination change, blur, incompleteness and noise etc.).

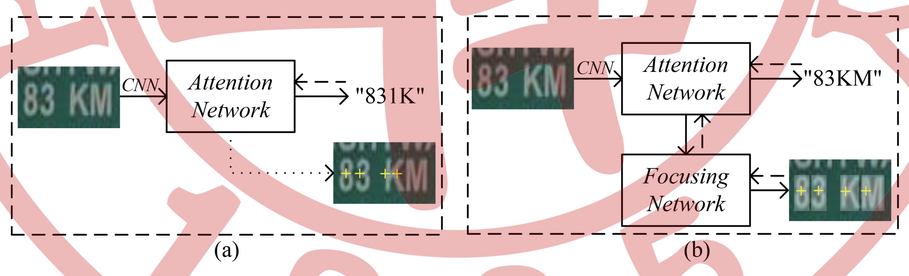


The "attention drift" in the attention-based model

- The existing attention-based methods perform poorly on complicated or low-quality images.
- One major reason is the alignments estimated by the attention model are easily corrupted due to the complexity or low-quality of images.
- The attention model cannot accurately associate each feature vector with the corresponding target region in the input image. We call this phenomenon ***attention drift***. (the following figure (a))

Accurately recognizing texts with Focusing Attention

- We propose a novel method called FAN (the abbreviation of Focusing Attention Network) to accurately recognize text from natural images.
- The FAN is made of two major subnetworks: an attention network for character generation and a focusing network for automatically adjusting the attention region. (the following figure (b))

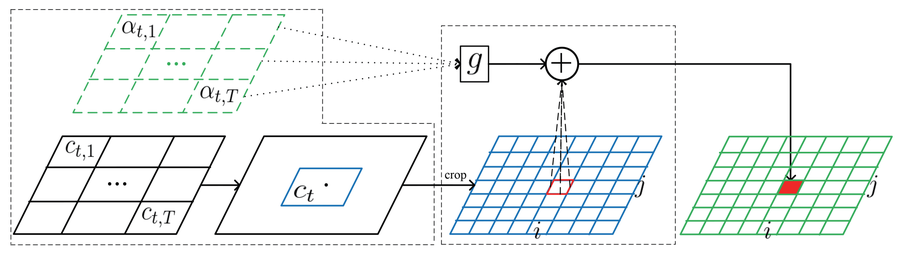


Materials and Methods

Datasets

Datasets	Description
SVT	collected from the Google Street View, consisting of 647 word images in its test set. Each image is associated with a 50-word lexicon.
ICDAR 2003	containing 867 cropped images. The lexicons include the 50-word lexicons and the full lexicon which combines all lexicon words.
ICDAR 2013	the successor of IC03, from which most of its data are inherited. It contains 1015 cropped text images. No lexicon is associated.
ICDAR 2015	containing 2077 cropped images. For fair comparison, we discard the images that contain non-alphanumeric characters, which results in 1811 images. No lexicon is associated.
IIIT5K	containing 3000 cropped word images in its test set. Each image specifies a 50-word lexicon and a 1k-word lexicon.

The FAN Methods



- 1) **Computing attention center**
  - For position (x, y) in layer L, we compute its receptive field in layer L-1 as:

$$x_{min} = (x-1) \times stride_w + 1 - pad_w,$$
$$x_{max} = (x-1) \times stride_w - pad_w + kernel_w,$$
$$y_{min} = (y-1) \times stride_h + 1 - pad_h,$$
$$y_{max} = (y-1) \times stride_h - pad_h + kernel_h.$$

  - Computing the receptive field of h_j in the input image:

$$c_{t,j} = location(j)$$

  - The attention center of target y_t in the input image is :

$$c_t = \sum_j^T \alpha_{t,j} c_{t,j}$$

- 2) **Focusing attention over target regions**
  - Cropping a patch of feature maps from the input image or a convolution output:

$$\mathcal{F}_t = Crop(\mathcal{F}, c_t, \mathcal{P}_h, \mathcal{P}_w)$$

  - Computing energy distribution over the attention region:

$$e_t^{(i,j)} = \tanh(Rg_t + S\mathcal{F}_t^{(i,j)} + b)$$

  - The probability distribution over the selected region:

$$P_t^{(i,j,k)} = \frac{exp(e_t^{(i,j,k)})}{\sum_c^K exp(e_t^{(i,j,c)})}$$

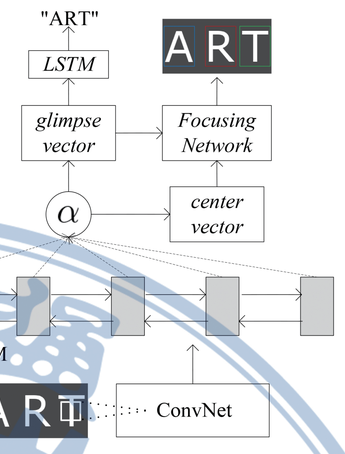
  - Focusing loss function:

$$\mathcal{L}_{focus} = -\sum_t^T \sum_i^{\mathcal{P}_w} \sum_j^{\mathcal{P}_h} logP(\hat{y}_t^{(i,j)} | \mathcal{I}, \omega)$$

The Network Training

- We combine a ResNet-based feature extractor, AN and FN into one network, as shown in right figure.
- The ResNet-based CNN architecture is as follows:

layer name	32 layers	output size
conv1_x	$3\times3, 1\times1, 1\times1, 32$ $3\times3, 1\times1, 1\times1, 64$	32×256
	pool:$2\times2, 2\times2, 0\times0$	
conv2_x	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times1$	16×128
	$3\times3, 1\times1, 1\times1, 128$	
	pool:$2\times2, 2\times2, 0\times0$	
conv3_x	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	8×64
	$3\times3, 1\times1, 1\times1, 256$	
	pool:$2\times2, 1\times2, 1\times0$	
conv4_x	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times5$	4×65
	$3\times3, 1\times1, 1\times1, 512$	
conv5_x	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times3$	1×65
	$2\times2, 1\times2, 1\times0, 512$	
	$2\times2, 1\times1, 0\times0, 512$	



Results and Conclusions

Results on several benchmarks

Method	IIIT5k			SVT		IC03			IC13	IC15
	50	1k	None	50	None	50	Full	None	None	None
ABBYY [29]	24.3	–	–	35.0	–	56.0	55.0	–	–	–
Wang *et al.* [29]	–	–	–	57.0	–	76.0	62.0	–	–	–
Mishra *et al.* [8]	64.1	57.5	–	73.2	–	81.8	67.8	–	–	–
Wang *et al.* [31]	–	–	–	70.0	–	90.0	84.0	–	–	–
Goel *et al.* [6]	–	–	–	77.3	–	89.7	–	–	–	–
Bissacco *et al.* [4]	–	–	–	90.4	78.0	–	–	–	87.6	–
Alsharif and Pineau [2]	–	–	–	74.3	–	93.1	88.6	–	–	–
Almazán *et al.* [1]	91.2	82.1	–	89.2	–	–	–	–	–	–
Yao *et al.* [32]	80.2	69.3	–	75.9	–	88.5	80.3	–	–	–
Rodríguez-Serrano *et al.* [22]	76.1	57.4	–	70.0	–	–	–	–	–	–
Jaderberg *et al.* [12]	–	–	–	86.1	–	96.2	91.5	–	–	–
Su and Lu [26]	–	–	–	83.0	–	92.0	82.0	–	–	–
Gordo [7]	93.3	86.6	–	91.8	–	–	–	–	–	–
Jaderberg *et al.* [13]	97.1	92.7	–	95.4	80.7	98.7	**98.6**	93.1	90.8	–
Jaderberg *et al.* [12]	95.5	89.6	–	93.2	71.7	97.8	97.0	89.6	81.8	–
Shi *et al.* [24]	97.6	94.4	78.2	96.4	80.8	98.7	97.6	89.4	86.7	–
Shi *et al.* [25]	96.2	93.8	81.9	95.5	81.9	98.3	96.2	90.1	88.6	–
Baidu IDL. [27]	–	–	–	–	–	–	–	–	89.9	**77.0**
Baseline	98.9	96.8	83.7	95.7	82.2	98.5	96.7	91.5	89.4	63.3
FAN	**99.3**	**97.5**	**87.4**	**97.1**	**85.9**	**99.2**	97.3	**94.2**	**93.3**	70.6

- We can see that FAN achieves better performance than the baseline in all cases, and substantially outperforms the 18 existing methods on almost all benchmarks.
- Right picture shows some real images processed by AN and FAN. Comparing with the outputs of AN, FAN obviously rectify the attention drift problem and correctly recognize more characters in the images.:

Conclusions

- In this work, we put forward the attention drift concept to explain the poor performance of existing AN based methods of scene text recognition on complicated and/or low-quality images, and propose a novel method FAN to solve this problem.

Before **After**