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 IIT5k, 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 low-quality images.
- One major reason is that the alignments estimated by the attention model are largely corrupted due to the complexity and 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 texts from natural images.
- The FAN is made of two major sub-networks: an attention network for character generation and a focusing network for automatically adjusting the attention region. (the above image figure (b))

Materials and Methods

Datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVT</td>
<td>collected from the Google Street View, consisting of 647 word images in its test set. Each image is associated with a 50-word lexicon.</td>
</tr>
<tr>
<td>ICDAR 2003</td>
<td>containing 667 cropped images. The lexicon include the 50-word lexicon and the full lexicon which contains all lexicon words.</td>
</tr>
<tr>
<td>ICDAR 2013</td>
<td>the successor of ICDAR, from which most of its data are inherited. It contains 1055 cropped test images. No lexicon is associated.</td>
</tr>
<tr>
<td>ICDAR 2015</td>
<td>containing 3037 cropped images. For fair comparison, we discard the images that contain non- alphanumeric characters, which results in 1811 images. No lexicon is associated.</td>
</tr>
<tr>
<td>IIT5K</td>
<td>containing 3000 cropped word images in its test set. Each image specifies a 50-word lexicon and a 1k-word lexicon.</td>
</tr>
</tbody>
</table>