Single Shot Text Detector with Regional Attention
Pan He1, Weilin Huang2, Tong He3, Qile Zhu1, Yu Qiao3 and Andy Li1
1 National Science Foundation Center for Big Learning, University of Florida
2 Department of Engineering Science, University of Oxford
3 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Introduction

- **Goal:**
  - Improve the speed and accuracy for scene text detection.
- **Existing works:**
  - Pixel-based detectors [1]: Using cascaded Fully Convolutional Networks (FCN) to cast character-based detections into the problem of pixel-wise text semantic segmentation.
  - Box-based detectors [2]: Extending object detectors such as Faster R-CNN [3] or SSD [4] to predict text boxes by simply using bounding-box annotation.
- **Problem & Motivation:**
  - In spite of effectively identifying rough text regions, pixel-based text detectors fail to produce accurate word-level predictions with a single model. The main challenge is to precisely identify individual words from a detected rough region of text.
  - Box-based text detectors are often trained by simply using bounding-box annotations, which may be too coarse (high-level) to provide a direct and detailed supervision, compared to the pixel-based detectors where a text mask is provided.
- **Our idea:**
  - We proposed techniques to bridge the gap between the pixel-based detectors and the box-based detectors, resulting in a single-shot model that essentially works in a coarse-to-fine manner.

Framework of SSTD with Regional Attention

![Framework of SSTD with Regional Attention](image)

Text Attention Module

- **Idea:**
  - A top-down spatial attention on text regions to suppress the background interference and cast the cascaded FCNs detectors into a single model.
- **Attention Map:**
  - We compute a text attention map from Aggregated Inception Features (AIFs).
  - The attention map indicates rough text regions and is further encoded into the AIFs (via element-wise dot production).
  - The attention module is trained by using a pixel-wise binary mask of text.

Hierarchical Inception Module

- **Idea:**
  - Aggregating inception features in different layers (with varied resolutions) to enhance local detailed information and encode richer context information.
  - **Aggregated Inception Features:**
    - Similar to Inception architecture in GoogLeNet [5], we get inception features through different convolutional operations, with Aggregated Inception Features.
    - We further enhance the inception features by aggregating multi-layer inception features, by using channel concatenation.
    - Each AIF is computed by fusing the inception features of current layer with two directly adjacent layers.

Experiment Results

- **Comparisons with state-of-the-art results:**
  - Table 1: Performances on the ICDAR 2013 and ICDAR 2015 datasets
  - Table 2: Performance on COCO-text dataset
  - Table 3: Exploration study on the ICDAR 2013 dataset
  - Table 4: Qualitative results:

Reference