Character Region Awareness for Text Detection

Younmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee

Clova AI Research, NAVER corp.

1. Abstract

- We propose a new scene text detection method to effectively detect text area by exploring each character and affinity between characters.
- Our proposed framework exploits both the given character-level annotations for synthetic images and the estimated character-level ground-truths for real images acquired by the learned interim model.

2. Introduction

- Unfortunately, most of the existing text datasets do not provide character-level annotations, and the work needed to obtain character-level ground truth is too costly.
- Our framework is designed with a convolutional neural network producing the character region score and affinity score.
    - *region score*: localize individual characters in the image
    - *affinity score*: group each character into a single instance
- To compensate for the lack of character-level annotations, we propose a weakly-supervised learning framework that estimates character-level ground truths in existing real word-level datasets.
- Visualization of character-level detection using CRAFT.

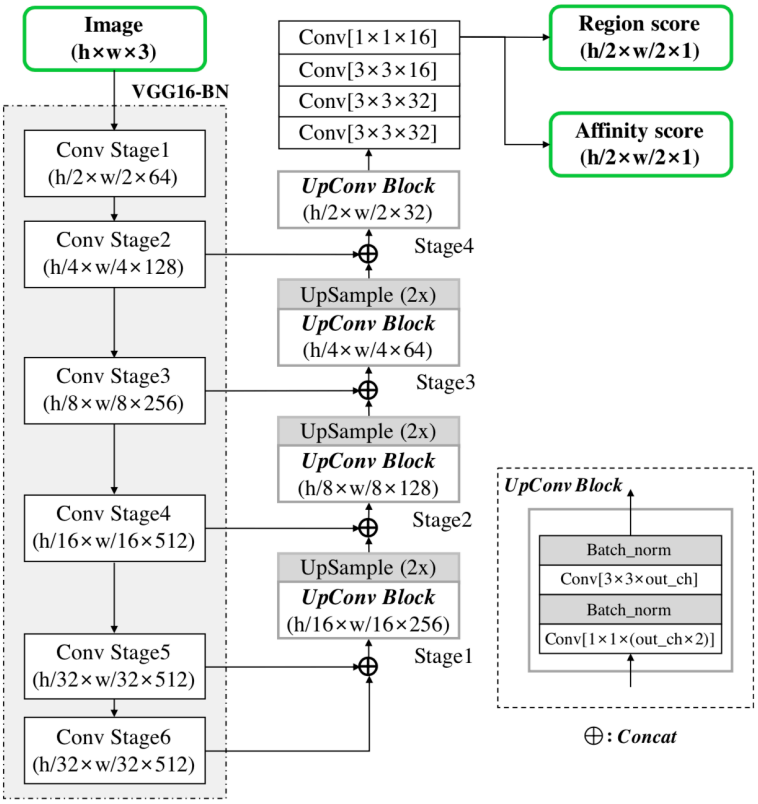
(a) (b)

3. Methodology

- Our main objective is to precisely localize each individual character in natural images.

3.1 Architecture

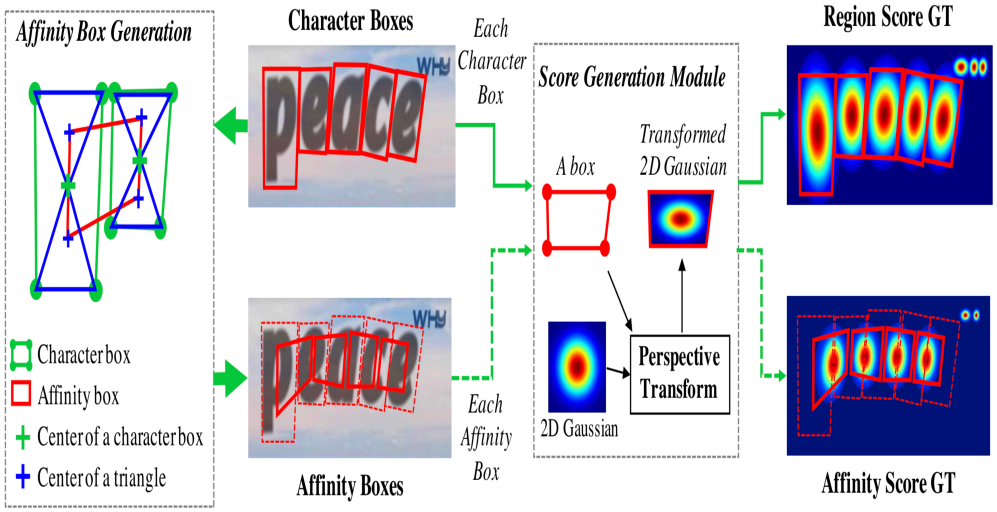
- backbone : VGG-16 (with batch normalization) + skip connections (in decoding part)
- output : *region score* and *affinity score*



3.2 Training

3.2.1 Ground Truth Label Generation

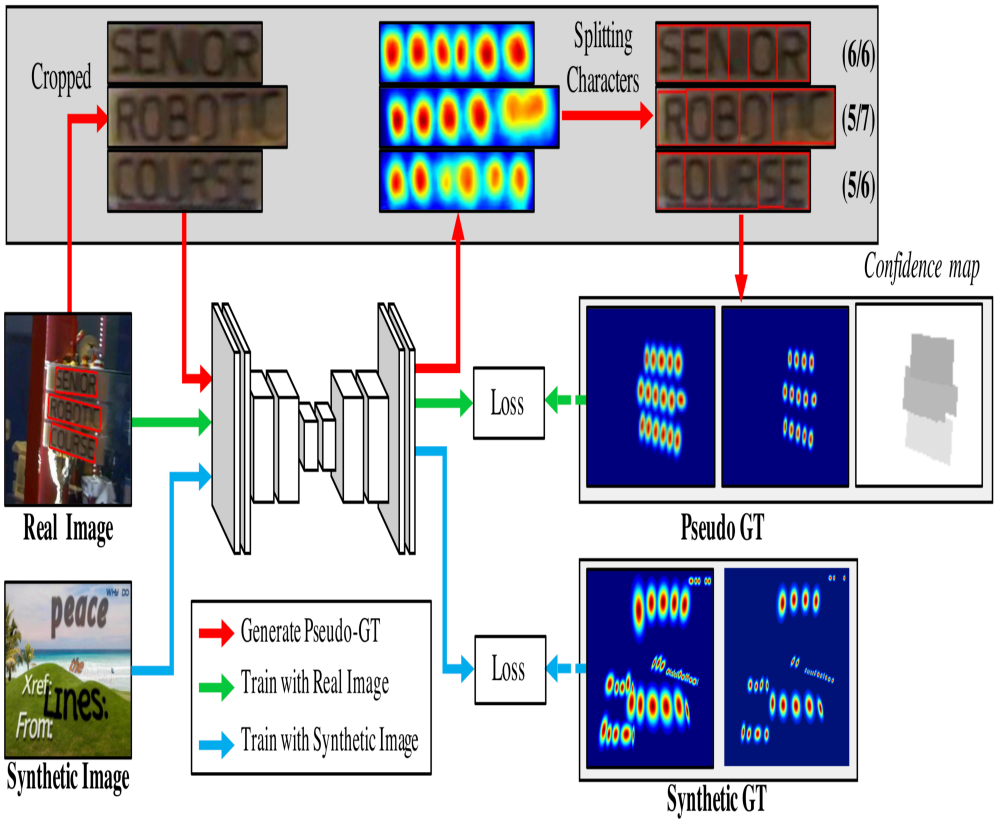
- Generate the ground truth label for the region score and the affinity score with character-level bounding boxes.
  - *region score* : probability that the given pixel is the center of the character
  - *affinity score* : center probability of the space between adjacent characters
- Unlike a binary segmentation, we encode the probability of the character center with a Gaussian heatmap.



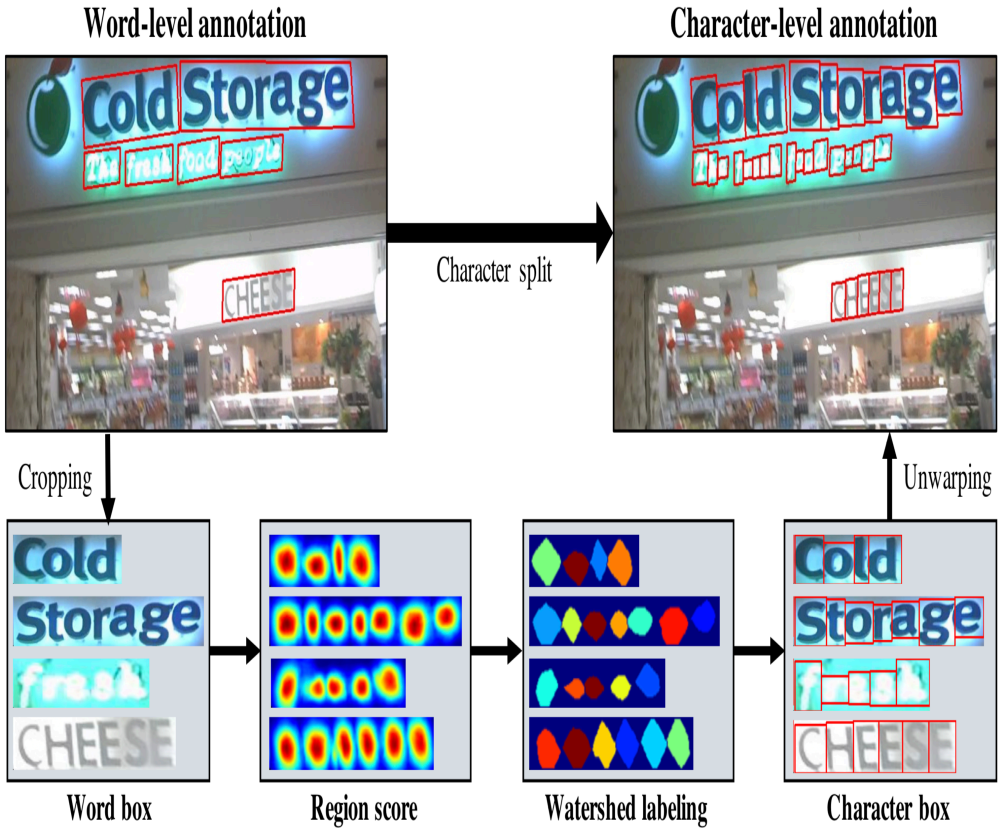
- Computing the Gaussian distribution value disrectly for each pixel within the bounding box is very time consuming.

- Since character bounding boxes on an image are generally distorted, we use the following steps.
    1. prepare 2D Gaussian map. (isotropic)
    2. compute perspective transform between the Gaussian map and character box
    3. warp Gaussian map to the box area.
- The proposed ground truth definition enables the model to detect large or long-length text instances, despite using small receptive fields.

- **3.2.2 Weakly-Supervised Learning**
    - When a real image with word-level annotations is provided, the learned interim model predicts the character region score of the cropped word images to generate character-level bounding boxes.
    - Illustartion of the overall training stream for the proposed method.



    1. Word-level images are cropped from the original image.
    2. Model trained up to date predicts the regions score.
    3. Character regions are splited by using Watershed algorithm
    4. Coordinates of the character boxes are transformed backed into the original image coordinates using the inverse transform from the cropping step.
    - Character split procedure for achieving character-level annotation from word-level annotation



- Measuring the quality of each pseudo-GTs generated by the model.
  - Confidence score.

$$s_{conf}(w) = \frac{l(w) - min(l(w), |l(w) - l^c(w)|)}{l(w)}$$

$R(w)$: bounding box region of sample w

$l(w)$: word length w

$l^c(w)$: estimated character bounding boxes and their corresponding length of characters

  - Pixel-wise confidence map

$$S_c(p) = \begin{cases} s_{conf}(w) & p \in R(w) \\ 1 & otherwise, \end{cases}$$

p : the pixel in the region $R(w)$

  - Objective

$$L = \sum_p S_c(p) \cdot (\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2)$$

$S_r^*(p)$, $S_a^*(p)$: pseudo-ground truth *region score* and *affinity map*

$S_r(p)$, $S_a(p)$: predicted *region score* and *affinity score*

When training with synthetic data, we can obtain the real ground truth, so $S_c(p)$ is set to 1

- Character region score maps during training