EAST Preliminary Summary

Shishir Jakati
University of Massachusetts, Amherst
sjakati@umass.edu

Abstract

Scene text detection is a difficult problem due to the various orientations and sizes of text instances. EAST is able to advance progress on the problem by utilizing a U-Shaped architecture to concatenate a larger amount of lower and higher level features. In addition, EAST proposes a network output of two robust separate geometries to be used in detection.

1. Network Structure

1.1. U-Shaped Model

The detector model which EAST[4] utilizes is a U-shaped network, which mimics the network structure presented in U-Net[2]. Due to the nature of the task of scene text detection, detector networks must gather information from both higher-level and lower-level layers. The U-Net architecture, as utilized by EAST, contains a main feature detector stem that is constantly constricting. In a contrasting manner, the fully convolutional network is augmented by up sampling layers which increase the resolution of outputs from the constricted layers.

This underlying network architecture allows the network to use varying levels of image features while not increasing the computational overhead. Experiments were performed with both PVANet[1] and VGG16[3] as the main feature extractor stem, pretrained on the ImageNet dataset.

1.2. Label Generation

EAST, as a scene text detector, acts as a regressor over both textboxes, rotated textboxes, and quadrangles. As seen in the network diagram structure presented in the original publication, the authors of EAST incrementally formed textbox outputs by way of adding channels to each output respectively.

Axis Aligned Bounding Boxes (AABB), are used in the creation of Rotated Boxes. They are defined as such

$$AABB = R = \{d_i | i \in \{1, 2, 3, 4\})\}; RBOX = \{R, \theta\}$$

It is easy to see that the AABB and RBOX are represented by 4 and 5 channels respectively. The final outputted geometry is a quadrangle, QUAD. These are defined as such

$$QUAD = Q = \{(\Delta x_i, \Delta y_i) | i \in \{1, 2, 3, 4\}\}$$

Here each pain $(\Delta x_i, \Delta y_i)$ represents the coordinate shift from the top left pixel location. Since there are 4 such points, there are a total of 8 channels in each **Q**.

1.3. Loss Functions

The loss is defined as

$$L = L_s + \lambda_g L_g$$

where L_s is the score map loss, and L_g is the geometry loss. In this case the λ constant represents the importance between the two losses.

1.3.1 Score Map

The loss on the score map is simply a balanced cross-entropy term. This is used to mitigate any bias the network may develop in terms of negative classification of text regions. The loss term is defined as follows

$$L_s = -\beta Y^* log(\hat{Y}) - (1 - \beta)(1 - Y^*) log(1 - \hat{Y})$$

where

$$\beta = 1 - \frac{\sum_{y^* \in Y^*} y^*}{|Y^*|}$$

In this case, $\hat{Y}$ are the predicted text/non-text labels, and Y^* are the ground truth text/non-text labels.

1.3.2 Geometries

The loss on geometries is more involved. Due to the nature of scene text, the loss function itself my balance large and small text instances. For this reason, the geometry loss is defined to be scale-invariant. Additionally, both the rotated textboxes and quadrangles have separate losses.

Rotated Textbox The rotated textbox loss is defined as follows

$$L_g = L_{AABB} + \lambda_\theta L_\theta$$

Intuitively, the loss function is composed of the loss on axis-aligned bounding boxes and the loss on the angle of rotation. By practical consideration, there is a constant to maintain how important the rotation loss is, this was set to one in the original experiments. The loss on the axis-aligned bounding boxes is simply a negative log functin of the Intersection Over Union (IoU) of box predictions. It is defined as follows

$$L_{AABB} = -log(|\frac{|\hat{R} \cap R^*|}{|\hat{R} \cup R^*|})$$

Here R^* is the predicted AABB geometry and $\hat{R}$ is the corresponding ground-truth label. The specific implementation of the IoU on bounding boxes is that the width and height of the intersected rectangle is computed and used as the numerator of the apparent fraction, while the sum of the two bounding boxes respectively less the intersection is used as the denominator.

The loss on rotation angles is quite simple, and is defined as follows

$$L_\theta = 1 - cos(\hat{\theta} - \theta^*)$$

Here $\hat{\theta}$ is simply the predicted angle of rotation, and θ^* is the corresponding ground truth rotation angle.

Quadrangles The loss on quadrangles is simply a smoothed L1 term with a normalization constant which accounts for the fact that text detections generally have one side with a dominating length. It is defined as follows

$$L_g = L_{QUAD}(\hat{Q}, Q^*)$$

and

$$L_{QUAD} = \min_{\hat{Q} \in P_{Q^*}} \sum_{c_i \in C_Q, \tilde{c_i} \in C_{\tilde{Q}}} \frac{smoothed_{L1}(c_i - \tilde{c_i})}{8 \times N_{Q^*}}$$

Here, N_{Q^*} is the normalization term, which is the shorted edge length of the quadrangle. C_Q is the ordered set of coordinates of some quadrangle Q.

1.4. Locality Aware NMS

A locality based NMS algorithm is proposed in the original publication. It is not reproduced here, though the idea is to reduce the number of candidate geometries for supression to a reasonable number. This is done by a row-by-row merge operation on geometries.

2. Relevance to Research

The proposed framework is heavily engineered for scene text detection. The loss functions take into account the practical considerations associated with problems in scene text detection. Using balanced cross entropy is a common theme shared between multiple scene text detection schemes, thus this technique may be useful in training a sufficient detector when image resolution is less than ideal. Having smaller detectable regions would exponentially increase the ratio of negative pixels to positive pixels, with respect to text instances.

Another seemingly relevant aspect of the publication is the idea of a U-Shaped architecture. As U-Net was used to derive both lower level and higher level features from pixel data, a low resolution detector may be able to generate specific contexts from higher level features as well as specifics text information from lower level features.

References

[1] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. Pvanet: Deep but lightweight neural networks for real-time object detection, 2016. 1

[2] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 1

[4] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector, 2017. 1