An Attention-based Text Detection and Recognition Method for Terminals of Current Transformer’s Secondary Circuit

Da Lu¹*, Jia Liu¹ and Helong Li¹

¹ China Electric Power Research Institution, Beijing, 100192, China
*Corresponding author’s e-mail: luda@epri.sgcc.com.cn

Abstract. Recognizing irregular text in real industrial scenes is a challenging task due to the background clutter, low resolutions or distortions. In this work, an attention-based text detection and recognition method for terminals of current transformer’s secondary circuit is proposed. It consists of three major components: pre-processing, text detection and text recognition. In text recognition module, a novel spatial temporal embedding is designed to better utilize the positional information. During training, the proposed framework only requires sequence-level annotations, instead of extra fine-grained character-level boxes or segmentation masks as in previous work. Despite its simplicity, the proposed method achieves good performance on the dataset collected in actual working scene.

1. Introduction

With the speedy development of our society, the power system has begun to play an increasingly important role in our daily lives. As one of the key components of the substation, current transformer’s secondary circuit has a good many problems, such as large number of equipment screen cabinets and various types of external wiring. These problems bring about results of high maintenance complexity and labour consuming. With the proposal of smart grid in recent decades, the concept of intellectualization has gradually become the developing goal of substations. Therefore, designing an automatic text detection and recognition method for terminals of current transformer’s secondary circuit is the current research hotspot.

In recent years, algorithms based on deep learning method strategies have achieved splendid results in both text detection and text recognition fields. As for text detection field, a great many previous works regard text areas as objects and adopt the framework of general object detection, such as Faster R-CNN [1], SSD [2]. They apply multiple branches to predict both the offsets and scores of bounding boxes. In order to address the problem of arbitrary orientation happening in text detection, RRPN [3] proposed inclined anchors with angle information and rotated RoI pooling layer. PMTD [4] improves Mask R-CNN [5] by performing soft mask segmentation and adding a plane clustering module at the end of network. While for text recognition algorithm, the general idea is extracting image features by CNN and then obtaining the sequence of characters by LSTM, e.g. CRNN [6] and RobustScanner [7]. So far, many scholars have tried to apply existed text detection and recognition methods to the field of terminals of current transformer’s secondary circuit. Yuan et al. proposed a fault diagnosis method for terminals of current transformer’s secondary circuit [8]. Zhou et al. introduce an intelligent identification system for equipment screen cabinet line label of substation secondary system [9]. However, it adopted EAST [10] as its text recognition method, whose recognition results still need to be improved.
In order to tackle the difficulty of text detection and recognition for terminals of current transformer’s secondary circuit, we propose an attention-based algorithm. The main contributions of this paper can be summarized as follows:

1. We present a complete framework of text detection and recognition process which can obtain location and content of all texts in the input image.
2. A novel spatial temporal embedding (STE) is designed for the attention-based text recognition module.

2. Methodology

The overall pipeline of our proposed method is shown in Figure 1. Specifically, our framework consists of three major components: pre-processing module, text detection module and text recognition module. In pre-processing module, the input figure will be resized to a fixed scale and accompanied by random data augmentation. Text detection module aims to detect the location of texts effectively, while the goal of text recognition module is to recognize characters in the detected areas. Components will be detailed in the following sections.

2.1. Text Detection Module

Inspired by Mask R-CNN, we adopt Pyramid Mask Text Detector as our text detection module. As shown in Figure 2, the backbone network is followed by two branches: bounding box (Bbox) branch and mask branch, which will firstly predict bounding boxes and then predict soft text mask for each bounding box. Predictions are then fed into the plane clustering algorithm to obtain refined coordinates of each quadrilateral text area.

2.2. Text Recognition Module

The text recognition module is an encoder-decoder architecture. Specifically, the encoder, which consists of one deep convolutional neural network and one LSTM-based sequence module, is aimed to map the input image to a fixed-length holistic feature. The decoder is another sequence module, which recovers the character sequence step by step based on 2D attention.
2.2.1. Encoder

As shown in Figure 3, the input image is firstly fed into a deep convolutional neural network, which results in a 2D feature map $F$ of size $H \times W \times D$, where $D$ is the number of channels. Note that $F$ is not only the input of the LSTM encoder, but also the context for the 2D attention network in decoding procedure. To better utilize the sequential information, we then feed the feature map to a LSTM model with 512 hidden size per layer. At each time step, the LSTM encoder receives one column of the 2D feature map followed by max-pooling along the vertical axis, and updates its hidden state $h_t$. After $W$ steps, we get the final hidden state $h_W$, which is regarded as the holistic representation of the input image and provided for decoding.

2.2.2. Decoder

The decoder is another LSTM model with the same architecture as the encoder LSTM model. During decoding, at each time step $t$, the decoder LSTM model first generate a hidden state $h_t$ based on last hidden state $h_{t-1}$ and the current input $x_t$,

$$h_t = \text{LSTM}(h_{t-1}, x_t), \quad x_t = \begin{cases} \text{<start>}, & \text{if } t = 0 \\ y_{t-1}, & \text{if } t > 0 \end{cases},$$

where $\text{<start>}$ is a special token standing for the start of the character sequence, and $y_{t-1}$ is the output at last decoding step. Some works regard $h_t$ as a query, each feature vector of $F$ as a key, to compute an attention map

$$A_{ij} = \text{softmax}(h_t^T \cdot F_{ij}),$$

where $F_{ij}$ represents the feature vector at position $(i, j)$ of $F$. This way of computing attentions actually ignores position information. Intuitively, we hope our model can attend to different position of $F$ at different time steps. For example, at the early stage of decoding, the left part of $F$ is more important.

Inspired by the positional encoding in Transformer [11], we design a novel spatial-temporal embedding (STE) for the decoder. Specifically, we add a spatial embedding (SE) to $F$, a temporal embedding (TE) to $h_t$, followed by linear transformations:

$$F_{ij}' = W_s \cdot (F_{ij} + SE_{ij}) + b_s,$$

$$h_t' = W_t \cdot (h_t + TE_t) + b_t.$$

Next, these two position-augmented representations are concatenated and used to regress the attention map with a regression layer:

$$A_{ij}' = \text{softmax}(W_{att} \cdot \text{concat}(F_{ij}', h_t') + b_{att}).$$

The glimpse vector is then computed as a weighted sum of feature map $F$,

$$g_t = \sum_{ij} A_{ij}' \cdot F_{ij}.$$

Finally, we use a linear layer with parameters $(W_{cls}, b_{cls})$ to classify the glimpse vector into one character:

$$y_t = \text{softmax}(W_{cls} \cdot g_t + b_{cls}).$$
The decoding procedure is repeated until $y_t$ becomes the special token <EOS>.

3. Experiments

3.1. Dataset and Evaluation Metrics
We evaluate the proposed framework on a dataset consisting of 502 images with 5120 manually annotated text instances. 250 of the images are randomly chosen for training and the rest for validation. For text detection task, performance is measured in terms of bounding box level precision, recall and F-score. For text recognition task, we simply use character sequence level accuracy as the evaluation metric.

3.2. Implementation details
The proposed method is implemented in PyTorch. Experiments are conducted on two Nvidia GeForce RTX2080 GPUs with 12 GB memory. For text detection model, we totally train 40 epochs, using SGD as the optimizer with batch size 64. The initial learning rate is 0.08 and decays to one-tenth of the previous at the 20th and 32th epoch. For text recognition model, we use the Adam optimizer with batch size 32 at training time, the learning rate is set to 0.001 initially, with a decay rate of 0.9 every 100 iterations until it reaches $10^{-5}$.

3.3. Results
We conducted comparative experiments on our dataset. Our detection and recognition results are presented in Table 1 and 2, respectively.

| Table 1. Text detection performance on our dataset |
|---------------------------------------------------|
| Method    | Precision | Recall | F-score |
|-----------|-----------|--------|---------|
| CTPN [12] | 0.754     | 0.643  | 0.694   |
| EAST      | 0.836     | 0.795  | 0.815   |
| RRPN      | 0.822     | 0.793  | 0.807   |
| Ours      | 0.914     | 0.873  | 0.893   |

As shown in Table 1, our method outperforms aforementioned models with the same ResNet50 backbone. According to the data given above, our framework leads to at least 7.8% improvements in terms of F-score.

| Table 2. Text recognition performance on our dataset |
|-----------------------------------------------------|
| Method            | STE module | Accuracy |
|-------------------|------------|----------|
| CRNN              | -          | 0.815    |
| RobustScanner     | -          | 0.892    |
| Ours              | -          | 0.901    |
| Ours              | √          | 0.920    |

Based upon the statistics shown in Table 2, our recognition method surpasses CRNN and RobustScanner by 10.5% and 2.8% in accuracy. Furthermore, we can see that STE module improves about 2% on our dataset, which proves the effectiveness of our proposed module.

4. Conclusion
In this paper, we propose an attention-based text detection and recognition method for terminals of current transformer’s secondary circuit, which can obtain texts location and characters automatically. The overall pipeline can be separated into three steps: pre-processing module, text detection module and text recognition module. Specifically, we adopt Pyramid Mask Text Detector as our text detection module and LSTM-based sequence module as our basic text recognition model. A novel spatial temporal embedding (STE) for the decoder is proposed to improve the ability of the network to capture important
information. Extensive experiments on dataset captured in real scene verify the effectiveness of our proposed method.

Acknowledgements
This work is supported by the Science and Technology Project of State Grid Corporation of China (Research and Application of On-site Inspection Technology Optimization on Gateway Electric Energy Metering Device, 5600-201926164A-0-0-00).

References
[1] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. J. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6): 1137–1149.
[2] Liu W., Anguelov D., Erhan D., et al. (2016) SSD: single shot multibox detector. In: European Conference on Computer Vision. Amsterdam. pp. 21–37.
[3] Ma, J., Shao, W., Ye, H., et al. (2018). Arbitrary-oriented scene text detection via rotation proposals. J. IEEE Transactions on Multimedia. 20: 3111-3122.
[4] Liu, J., Liu X., Sheng, J., et al. (2019) Pyramid Mask Text Detector. https://arxiv.org/abs/1903.11800.
[5] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017) Mask R-CNN. In: IEEE International Conference on Computer Vision. Venice. pp. 2980-2988.
[6] Shi, B., Bai, X., Yao, C. (2017) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. J. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(11): 2298-2304.
[7] Yue, X., Kuang, Z., Lin, C., et al. (2020) RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition. In: European Conference on Computer Vision. Virtual. pp. 135-151.
[8] Yuan, G.Q. (2020) Research on Secondary Circuit Detection Technology of Intelligent Substation. (Degree of Master, Guangdong University of Technology).
[9] Zhou, J., Li, Y., Zhang, K.S. (2018) Intelligent Identification System for Equipment Screen Cabinet Line Label of Substation Secondary System. Machinery and Electronics. Vol.36, No.11: 67-70.
[10] Zhou, X., Yao, C., Wen, H., et al. (2017) EAST: An Efficient and Accurate Scene Text Detector. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2642-2651.
[11] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention is all you need. In: Advances in Neural Information Processing Systems. California. pp. 6000-6010.
[12] Tian Z., Huang W., He T., et al. (2016) Detecting Text in Natural Image with Connectionist Text Proposal Network. In: European Conference on Computer Vision. Amsterdam. pp. 56–72.