A text detection algorithm for ancient Chinese writing based on Swin transformer and mask R-CNN

Sichen Hao, Youguang Chen*
School of Data Science and Engineering, East China Normal University, Shanghai, China

ABSTRACT

The study of ancient Chinese writing has great cultural and historical value. Text annotation is a time-consuming and labor-intensive part of ancient writing research. In this paper, we construct a deep learning model for ancient Chinese text detection which combines Swin Transformer and Mask R-CNN. To solve the overlapping detection problem, we propose Text Non-Maximum Suppression (text-NMS) and text-NMS loss. The former is to weed out redundant subtext bounding boxes in the Non-Maximum Suppression process, and the latter further rectifies the detection failure missed by text-NMS in bounding box regression. Experiments on the Chinese Stone Inscription Dataset show that the proposed algorithm can improve the accuracy of ancient text detection. The text-NMS and text-NMS loss algorithm boost the average precision (AP) of Swin-small and Mask R-CNN from 64.4% to 65.8% with few additional hyper-parameter and computational overhead.

Keywords: Deep learning, text detection, non-maximum suppression, Swin transformer, mask R-CNN

1. INTRODUCTION

Writing is one of the most influential creations of human civilization. As a country with a long history, China has a large number of ancient books and cultural relics. The ancient characters carried on them are essential materials for today's historical research. With the development of the times, researchers have begun digitalizing these materials in order to better preserve and facilitate future research. Text annotation is an essential part of digitalization. It used to be done manually, requiring corresponding professional knowledge and sufficient time and effort. Using computers to complete the detection and recognition task of ancient writing can effectively improve work efficiency and accuracy.

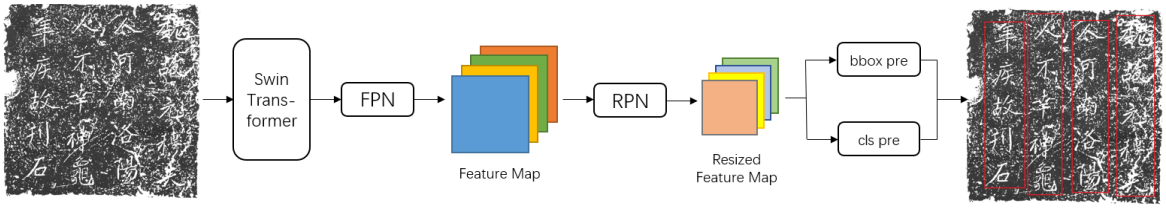

Figure 1. The procedure of ancient text detection.

Text detection is one of the object detection applications in which an algorithm predicts the location of text lines in an image with bounding boxes and gives a confidence score to the box. It is often used for Optical Character Recognition (OCR) along with text recognition. Traditional text detection algorithms use handcrafted features to detect the text[1,2]. These methods rely on artificially designed feature descriptor, which has a poor performance in detecting blurry texts or diverse handwriting. With the rapid development of deep learning algorithms, neural network-based feature extraction methods gradually replaced handcrafted features. Some scholars proposed algorithms relying on Convolutional Neural Network (CNN) which achieved good performance in modern text detection. However, ancient text pictures have the characteristics of serious background noise, different text scales, and complicated fonts, which differs from modern Chinese writing or other languages. There are still few studies about ancient Chinese text detection. In this paper, we proposed a method based on Swin Transformer[3] and Mask R-CNN[4]. The procedure of the algorithm is shown in Figure 1. The method takes advantage of Swin Transformer's powerful feature extraction ability on dense objects and uses Mask R-CNN to predict the location of text bounding boxes. The main contributions of this paper are as follows:

* ygchen@cc.ecnu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250602
 · 0277-786X · doi: 10.1117/12.2661859

(1) Inspired by soft-NMS[5], we put forward a new non-maximum suppression (NMS) algorithm for text detection called text-NMS, which can process overlapping bounding boxes more accurately.

(2) We improve the bounding box regression loss function of Mask R-CNN. A new loss function called TNMS-Loss (Text Non-Maximum Suppression Loss) is added to the original Smooth L1 Loss. The new loss term further solves the problem of overlapping detection based on text-NMS.

2. RELATED WORK

2.1 Feature extraction backbone

Based on the self-attention mechanism, Transformer was first applied to natural language process (NLP) tasks and achieved excellent performance[6]. Transformer's strong representation capability and ease of applicability made it popular in the NLP field. In 2020, scholars tried to apply transformer structure to Computer Vision (CV) tasks and found out that it also performed well in image feature extraction[7,8], showing its potential in replacing CNN as the vision backbone. Swin Transformer[3] was proposed by Liu et al. based on the work of ViT[8]. It introduced patch merging on attention windows to get features from low-level to high-level. And through the attempt of shifted window partitioning, connections among different windows were established to broaden the model's receptive field. Such a feature extraction method enables it to fit well with detector necks like Feature Pyramid Network (FPN)[9] to deal with multi-scale changes in object detection.

2.2 Object detection model

Object detection algorithms can be divided into two categories: one-stage detection like Reference[10] and two-stage detection like References[4,11]. One-stage detectors are more efficient, while two-stage detectors are more accurate. In this paper, we use the two-stage model Mask R-CNN[4] as the basic detector of ancient Chinese text detection. Mask R-CNN has been widely used in object detection and instance segmentation. It was proposed based on Faster R-CNN[11], which first used the Region Proposal Network (RPN) to give predictions of detection bounding boxes. Due to RPN, Faster R-CNN achieved real-time level detection much quicker than two-stage detectors before. In Mask R-CNN, the author added the mask prediction part to Faster R-CNN and replaced ROI Pooling in RPN with ROI Align, enabling the model to perform pixel-level segmentation.

2.3 Non-maximum suppression

NMS is used to pick out the best result from a set of candidate boxes describing the same target obtained by the detector. It utilizes the classification score of candidate boxes and Intersection over Union (IoU) between two boxes to filter redundant boxes. According to the traditional NMS algorithm, boxes with lower scores will be deleted if their IoU with higher score boxes exceeds the given threshold. Bodla et al.[5] put forward that the traditional NMS may remove some correct boxes if one object has a major overlap with the other one. Instead of deleting boxes "hardly", they give bounding boxes of lower scores a decayed score depending on a continuous function of IoU. The method was described as soft-NMS. It was proved to bring accuracy improvement to common detection tasks. But in the cases of text detection, especially ancient Chinese text detection, there are still some problems.

(a) (b) (c)

Figure 2. Examples of problems when applying the traditional NMS algorithm.

As shown in Figure 2, two overlapping boxes are both reserved in the detection results of ancient text rubbings when using traditional NMS. The smaller box is usually the subtext of the whole text. Such an issue was also encountered in soft-NMS. That's because the IoU threshold is not low enough to filter the smaller bounding box. But if we turn down the threshold, correct boxes will be deleted for a small amount of overlap with other right boxes, which may also lead to detection failure. A single IoU threshold is not enough to deal with all situations.

3. OUR PROPOSED METHOD

3.1 Text non-maximum suppression

As mentioned above, adjusting the IoU threshold cannot better solve the problem of overlapping detection. Aiming at this situation, we propose a non-maximum suppression algorithm, text-NMS, which is suitable for text detection. The pseudocode of the algorithm is shown in Figure 3.

In addition to filtering candidate boxes by the IoU between the two boxes, we added the ratio of overlapping area $inter(M, b_i)$ to the area of the candidate box $\mathbb{S}_M$ as the filter. We call it subtext suppression. Referring to the approach of soft-NMS, we give a penalty to the score of corresponding boxes instead of removing them to avoid excessive suppression. A continuous Gaussian penalty function is used to decay the detection score, which can be written as follows:

$$s_i = s_i e^{-\frac{1}{\sigma}\left(\frac{inter(\mathrm{M},b_i)}{\mathbb{S}_\mathrm{M}}\right)^2} \tag{1}$$

$$s_i = s_i e^{-\frac{1}{\sigma}iou(\mathrm{M},b_i)^2} \tag{2}$$

Some candidate boxes may exceed both thresholds at the same time. Considering that the suppression effect of the two thresholds is similar, the suppression which poses a greater decrease to the detection score will be performed in this case. The detection result can be optimized by adjusting the NMS threshold and subtext suppression threshold separately. For N detection boxes, the computational complexity for text-NMS is $O(N^2)$, which is the same as traditional NMS and soft-NMS.

Input: initial detection boxes set $B = \{b_1,..,b_N\}$
detection scores set $S = \{s_1,..,s_N\}$
NMS threshold N_t
sub-text suppression threshold N_S
Output: result detection boxes set D
corresponding detection scores set S

$D \leftarrow \{\}$
while $B \neq empty$ **do**
  $m \leftarrow \arg\max S$
  $M \leftarrow b_m$
  $D \leftarrow D \cup M; B \leftarrow B - M$
  for b_i *in* B **do**
    $s_1 \leftarrow 1; s_2 \leftarrow 1$
    if $inter(M, b_i)/\mathbb{S}_i \geq N_S$ **then**
      $s_1 \leftarrow s_i f(inter(M, b_i)/\mathbb{S}_i)$ Subtext Suppression
    end
    if $iou(M, b_i) \geq N_t$ **then**
      $s_2 \leftarrow s_i f(iou(M, b_i))$ Soft-NMS
    end
    $s \leftarrow min(s, s_1, s_2)$
  end
end
Return D, S

Figure 3. Pseudocode of text non-maximum suppression algorithm.

After applying text-NMS, most of the overlapping box issues are solved. As the result example shown in Figure 4, the score of the subtext box is reduced to a lower value, which can be filtered by the score threshold of the detector. Since text-NMS is a greedy algorithm like traditional NMS, the algorithm does not take effect when the score of subtext is higher

than the whole text. But the overlap area is relatively small for the larger box, so the score of the large box will not be reduced or reduced slightly according to equation (1).

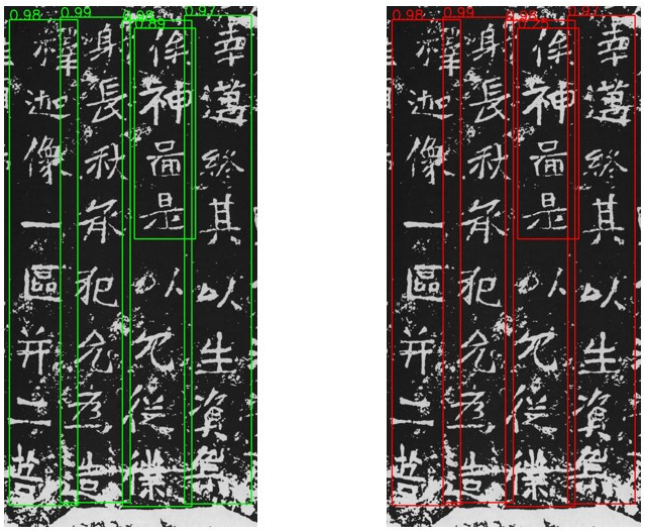

Figure 4. Example for NMS result. The left image with green bounding boxes is for traditional NMS, while the right one with red bounding boxes is for Text-NMS.

3.2 Additional loss term based on NMS

The loss function of Mask R-CNN can be expressed as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} \tag{3}$$

In the equation, Smooth L1 Loss is used for bounding box regression loss, which is written as below:

$$\mathcal{L}_{smoothL1} = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, |x| \geq 1 \end{cases} \tag{4}$$

However, the loss function only focuses on the distance of width, height, and center coordinate between the result and ground truth respectively, ignoring the correlation between coordinates. To solve this problem, Zheng et al. proposed CIoU Loss[12] which is defined as:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v \tag{5}$$

CIoU loss regards the 4-point box as a whole for regression and takes the aspect ratio into consideration which achieves faster convergence and better regression accuracy. Based on the CIoU loss and the proposed text-NMS algorithm, we added the Text-NMS loss term to the bounding box regression loss. It can be written as follows:

$$\mathcal{L}_{TNMS} = \begin{cases} \frac{inter(b_{gt},b_i)}{\mathbb{S}_i} - N_S & , \frac{inter(b_{gt},b_i)}{\mathbb{S}_i} > N_S, \frac{inter(b_{gt},b_i)}{\mathbb{S}_{gt}} < N_S \\ 0 & , others \end{cases} \tag{6}$$

In the equation, N_S is the subtext suppression threshold mentioned in text-NMS. b_i and b_{gt} stands for the predicted box and the ground truth box. The addition of the TNMS loss term strengthens the model's attention on the coincidence of the prediction and the ground truth. And as mentioned before, text-NMS cannot suppress subtext boxes whose scores are higher than the correct ones. The text-NMS loss can rectify these missed boxes in the regression. The total loss of bounding box regression we used is:

$$\mathcal{L} = \mathcal{L}_{CIoU} + \mathcal{L}_{TNMS} \tag{7}$$

With text-NMS and text-NMS loss, the situation of overlapping subtext detection failure can be basically avoided.

4. EXPERIMENTS

We perform experiments on the Chinese Stone Inscription Dataset using 2 TITAN RTX GPU.

4.1 Datasets

The dataset is collected from rubbings of stone inscriptions from the Wei Jin, Southern and Northern Dynasties (a period in China that lasted from 220 to 589 A.D.). All text is written vertically. There are 1776 images in the dataset, and these images contain 29767 text columns in total. 80% of images are used for training and the rest are for testing. The text regions are annotated by the coordinate of two diagonal vertices in COCO [13] format.

4.2 Results and analysis

We use Swin and Mask R-CNN as the baseline model. We adopt two versions of Swin Transformer in our experiments: Swin-Tiny and Swin-Small. "Tiny" and "Small" refer to the depth of the network and the number of parameters. With each backbone, two models are built to validate the performance of the proposed algorithms. The first one only replaces the traditional NMS with text-NMS, and the second one adds text-NMS loss term based on the first model. N_t and N_s are set to 0.5 and 0.7 respectively, and σ is set to 0.7 with the Gaussian penalty function. We use the following metrics to evaluate the performance of different models such as average precision (AP) @ 0.5, AP @ 0.5:0.95, and average recall (AR) @ 100. AP @ 0.5 is the area of the precision-recall curve when the IoU of prediction and ground truth larger than 0.5 is regarded as positive. And AP @ 0.5:0.95 is average value of AP from 0.5 to 0.95 in steps of 0.05. AR @ 100 stands for the average recall in the top 100 predicted boxes of each image.

Table 1 shows the results obtained from the experiments on the dataset. Our proposed method achieves 65.8% AP @ 0.5:0.95, 1.4% higher than the original Swin-Small with Mask R-CNN detector. It also gives rise to the AP @ 0.5 and AR @ 100 by 1.3% and 0.9%. And the algorithm gets similar results in the Swin-Tiny backbone. Some results detected by the best model are shown in Figure 5, the model performs well when the background is blurry and the writing is scribbled.

Table 1. Results on Chinese stone inscription dataset for our proposed method.

Backbone	Detector	AP @ 0.5	AP @ 0.5:0.95	AR @ 100
Swin-Tiny[3]	[4]	87.6%	60.5%	66.9%
Swin-Tiny	[4] + TNMS	88.3%	61.4%	67.4%
Swin-Tiny	[4] + TNMS + TNMS loss	88.8%	61.8%	67.6%
Swin-Small[3]	[4]	89.2%	64.4%	70.2%
Swin-Small	[4] + TNMS	89.9%	65.5%	70.9%
Swin-Small	[4] + TNMS + TNMS loss	90.5%	65.8%	71.1%

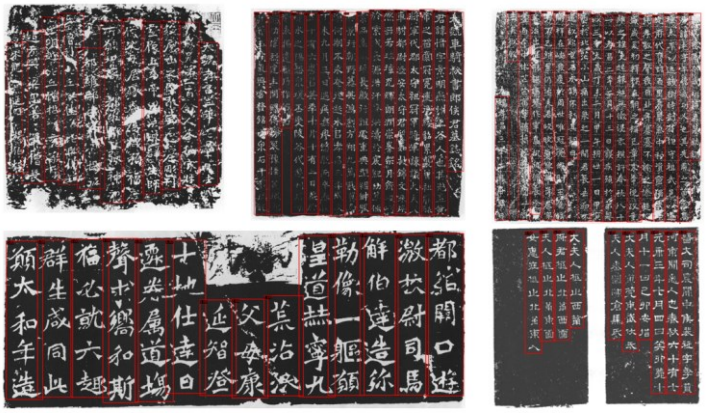

Figure 5. Several results of proposed model on different images.

5. CONCLUSIONS AND FURTHER WORK

In this paper, we construct an ancient Chinese text detection model with Swin Transformer as the feature extraction backbone and Mask R-CNN as the detector. In order to solve the overlapping detection problem, we proposed text-NMS in the non-maximum suppression stage and text-NMS loss in the regression stage. The proposed model shows good performance in Chinese ancient text detection, which can be applied to the study of palaeography. Since Mask R-CNN supports instance segmentation, which is not utilized in our work, we are complementing the segmentation annotation of the Chinese Stone Inscription Dataset for higher detection granularity. We are also trying to realize optical character recognition of ancient Chinese books by combining text detection and text recognition.

REFERENCES

[1] Epshtein, B., Ofek, E. and Wexler, Y., "Detecting text in natural scenes with stroke width transform," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2963-2970 (2010).

[2] Matas, J., Chum, O., Urban, M., et al., "Robust wide-baseline stereo from maximally stable extremal regions," Image and Vision Computing 22(10), 761-767 (2004).

[3] Liu, Z., Lin, Y., Cao, Y., et al., "Swin transformer: Hierarchical vision transformer using shifted windows," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 10012-10022 (2021).

[4] He, K., Gkioxari, G., Dollár, P., et al., "Mask R-CNN," Proc. of the IEEE Inter. Conf. on Computer Vision (ICCV), 2961-2969 (2017).

[5] Bodla, N., Singh, B., Chellappa, R., et al., "Soft-NMS—Improving object detection with one line of code," Proc. of the IEEE Inter. Conf. on Computer Vision (ICCV), 5561-5569 (2017).

[6] Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention is all you need," Advances in Neural Information Processing Systems (NIPS) 30, (2017).

[7] Carion, N., Massa, F., Synnaeve, G., et al., "End-to-end object detection with transformers," European Conf. on Computer Vision (ECCV), 213-229 (2020).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, (2020).

[9] Lin, T., Dollár, P., Girshick, R., et al., "Feature pyramid networks for object detection," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2117-2125 (2017).

[10] Redmon, J., Divvala, S., Girshick, R., et al., "You only look once: Unified, real-time object detection," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 779-788 (2016).

[11] Ren, S., He, K., Girshick, R., et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems (NIPS) 28, (2015).

[12] Zheng, Z., Wang, P., Liu, W., et al., "Distance-IoU loss: Faster and better learning for bounding box regression Proc. of the AAAI Conf. on Artificial Intelligence 34(07), 12993-13000 (2020).

[13] Lin, T., Maire, M., Belongie, S., et al., "Microsoft coco: Common objects in context," European Conf. on Computer Vision (ECCV), 740-755 (2014).