Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 3, 2022

Location of Scene Text Based on Yolov7

Yulong Chang[1,2], Youchan Zhu[1,2], Kaili Cui[1,2], Fujun Guan[1,2] and Zheng Li[1,2, *]

[1] Department of Computer, North China Electric Power University, Baoding 071000, China
[2] Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, North China Electric Power University, Baoding 071000, China

[*] **Corresponding author**: Zheng Li (Email: yeziperfect@163.com)

Abstract: Text is everywhere in daily life, and it carries rich and accurate information. Natural scene text detection technology can be widely used in various fields of life. Chinese is an important tool to carry culture. Therefore, it is of great significance to study natural scene Chinese text detection. For text location in Chinese scene, we use target detection method to train based on YOLOv7 model, and obtain effective detection results.

Keywords: Text detection; YOLOv7; Deep learning.

1. Introduction

Words exist in all aspects of our life, which helps us understand things and exchange ideas. Text in natural scenes also contains rich semantic information. Text is more easily concerned than other objects. Judd et al. [1] found that text is more easily concerned than other objects by analyzing the eye movement data of 1003 images. Let the computer find the text contained in the picture in the natural scene picture, and identify it, and make good use of the semantic information transmitted by the text in the picture has great development prospects.

The initial focus of character recognition is to recognize the scanned text of machine-printed documents, that is, optical character recognition (OCR) [2] With the in-depth study of OCR, it has reached a fairly high level of accuracy. However, the success of OCR system is limited to scanning the text in the document, and the recognition rate of scene text is not high. Recognizing text from scene images is more challenging than traditional OCR. The location of scene text is the premise of text recognition. The background of the scene image is sometimes uneven and highly chaotic, and the location of the text in the scene is not uniform and its changeable form becomes the main problem of the scene text detection. The text location method is realized by the target detection model. Traditional target detection methods use machine learning algorithms to determine target candidate regions through sliding windows of different scales,and extract target features in the region, such as histogram of oriented gradients(HOG)[3],scale invariant feature transform(SIFT),local binary pattern(LBP), haar like features(Haar),etc. According to the target detection method. Ren et al[6] proposed a method based on predefined bounding box, which mainly uses Faster-RCNN (Fast Region-based evolutionary neural network). Liu [7] proposed a classical target detection framework. This method first generates multiple predefined bounding boxes with each pixel as the center, then judges whether these bounding boxes contain complete text, and finally refines the bounding boxes containing complete text to fit the text boundary. The network structure of Text Boxes proposed by Liao [8] is consistent with the network structure of SSD. His solution is to change the detected object into text. Text Boxes adopts the full convolution network structure. By predicting the confidence of the text bounding box and the coordinate offset between the bounding box and the default bounding box (the pre-designed initial bounding box), the coordinate information of the word bounding box is directly output at multiple feature layers. The feature of Text Boxes is that it adjusts the proportion of the convolution kernel in the convolution neural network, and transforms the traditional square into a long rectangle that is more consistent with the character of the text, so that the receptive field becomes a long rectangle, making it easier for the network to extract the character features. In addition, Text Boxes also adjusted the aspect ratio of the predefined bounding box and used the predefined bounding box with a larger aspect ratio to adapt to the shape of the text and reduce the difficulty of the regression task. The recognition speed of the Text Boxes method based on the first stage is significantly improved, but the recognition accuracy is not high for text recognition under complex light background and word recognition with too large character spacing. Shi [9] proposed the SegLink method, which is also improved based on SSD. Its core idea is not to directly detect words or text lines, but to first detect local areas of words or text lines, and then connect these local areas to form a complete word or text line. It divides the text detection task into two subtasks: detecting the connection between text segments and predicting the connection between segments. Among them, fragments are rectangular bounding boxes with directions, which are predefined bounding boxes, and they cover part of words or text lines. The connection between fragments refers to whether two fragments belong to the same word or text line. The segments with connection relationship are merged to generate the bounding box of the corresponding text. The SegLink method is similar to the Text Boxes method, except that it is difficult to detect large text, because the connection is used to connect adjacent fragments, and cannot be used to detect lines of text with relatively long spacing. Ma [10] improved the predefined bounding box and ROI pooling layer on the basis of Faster-RCNN network. Because the direction of the scene text is arbitrary, and the detection method based on the predefined bounding box uses a horizontal rectangular bounding box. In order to make the predefined bounding box more coincident with the text, the predefined bounding box used by RRPN is angled, with 30 degrees as the interval, so as to obtain the bounding box covering 360 degrees. Since the bounding box is angled at this

time, the extracted RoI features are also angled rectangles. RRPN proposes a RRoI pooling method to pool such features, so as to obtain horizontal features and pass them into the RCNN network. The specific method is to divide the angled RoI features into $N \times N$ regions on average, where N is the edge length of the pooling feature, then max pooling feature for each 1×1 region, and finally map the horizontal $N \times N$ features. Thus, the RRPN network can be used to detect text lines in any direction.

Due to the lack of relevant research on scene Chinese detection, and the accuracy of location is not ideal. In order to locate the text information in the scene image, we use the target detection method. Using Yolov7 [11] target detection model, a lightweight real-time and efficient text detection method is proposed. This method mainly adjusts parameters and optimizes network framework based on YOLOv7 target detection model. For the problem of Chinese text detection in scene images, we train the network on the ReCts dataset. We hope to provide a real-time and efficient detection method for Chinese text location in the scene.

2. Organization of the Text

2.1. Overall structure

At present, within the range of 5~600 FPS, the speed and accuracy of YOLOv7 target detection model are very efficient and accurate. Figure 1 shows the model structure of YOLOv7. The model consists of three parts, namely input part, backbone feature extraction network part and head part. The input part uniformly converts the size of the image input to the model to 640×640×3. For the input image of the input part, the backbone feature extraction network carries out the operation and transmission between the network layers to extract the effective feature extraction of the image. The detection head part fuses and extracts the eigenvalues of the specific layer output, similar to the FPN structure, and finally obtains the result prediction through the Detect module.

Yolo7 backbone feature extraction network introduces ELAN structure and MP structure on the basis of Yolov5. Among them, ELAN structure is stacked by different convolution blocks, and does not change the width and height of input feature layer. It strengthens the interaction between each feature layer through expansion, random combination and splicing, and improves the learning ability of the model. The MP structure is composed of two paths, namely, a 3×3 convolution block path and a Maxpool path.The width and height of the input feature layer are compressed to enhance the feature fusion capability of the network. The detection head is composed of SPPCSPC module, deeper ELAN structure and RepConv module.Among them, the SPPCSPC module adds a "residual edge" on the basis of the SPP module and uses the idea of CSPLayer for reference, and stacks it with the feature layer output after Maxpool operation. RepConv module [12] is a re-parameterized convolution architecture. During model training, parallel 1×1 convolution branches are added for 3×3 convolution layers. During model deployment, the parameters of the branch are re-parameterized to the main branch, and the 3×3 main branch convolution output is taken. The reasoning speed of the model is improved by increasing the training cost.

2.2. ELAN and deepened ELAN

ELAN module is an efficient network structure. By controlling the shortest and longest gradient path, the network can learn more features and has stronger robustness. As shown in Figure 2, ELAN has two branches. The first branch changes the number of channels through a 1×1 convolution. The second branch is more complex. It first passes through a 1x1 convolution module to change the number of channels. Then, four 3x3 convolution modules are used for feature extraction. Finally, four features are superimposed to get the final feature extraction result.

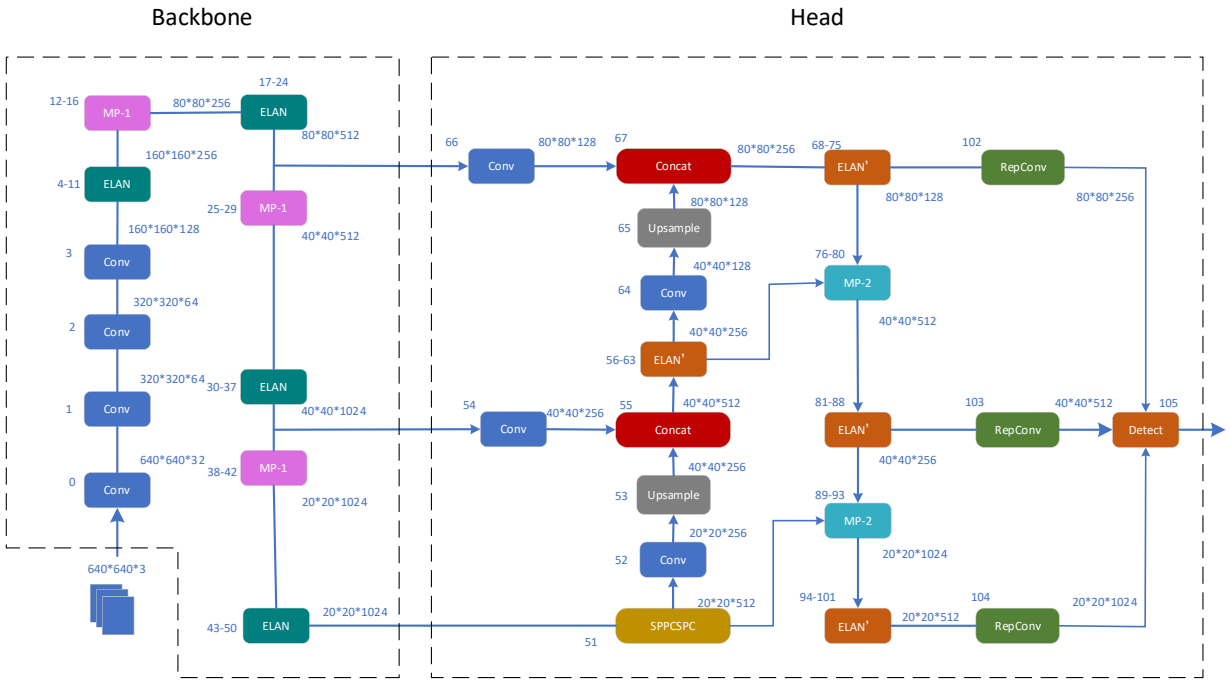
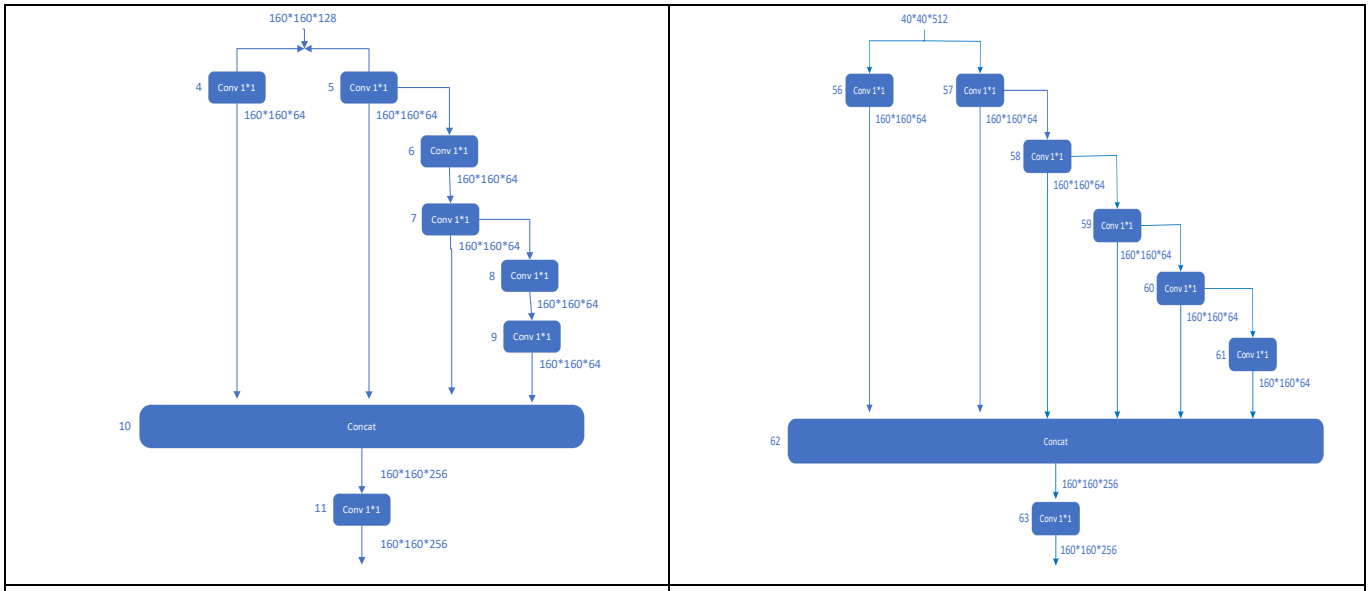


Figure 1. Yolov7 network structure diagram



a ELAN | b deeper ELAN

Figure 2. ELAN and ELAN ‘

The deepened ELAN, we can call ELAN'. For ELAN', it is very similar to the ELAN module. The slight difference is that the number of outputs it selects in the second branch is different. ELAN module selects three outputs for final addition. The ELAN'module selects five outputs to add.

2.3. SPPCSPC

The SPPCSPC structure is shown in Figure 3. First, the feature is divided into two different branches. The left part is processed like the SPP structure, and the right part is processed through convolution. Finally, the two parts are combined. In this way, the calculation amount can be reduced by half, making the speed faster and the accuracy improved. The left part can increase the receptive field to make the algorithm adapt to different resolution images. It obtains different receptive fields through a convolution layer. It can be seen from the figure that in the left branch, the eigenvalues pass through four branches with convolution layer in parallel. They are m [0], m [1], m [2] and the original eigenvalue without processing. These four different branches represent that they can process different objects.

2.4. MP module

The MP module can be divided into two structures according to the connected input. mp1 is shown in Figure 4a and mp2 is shown in Figure 4b. Mp module is divided into two branches, one of which is composed of Maxpool and convolution layer, and the other is composed of two convolution layers. Finally, the output results of the two branches are spliced. The main difference between Mp1 and Mp2 is that the input of Concat in Mp1 consists of two parts, with two branches respectively. The input of Concat in Mp2 consists of three parts, two of which come from two branches and the other from the backbone network.

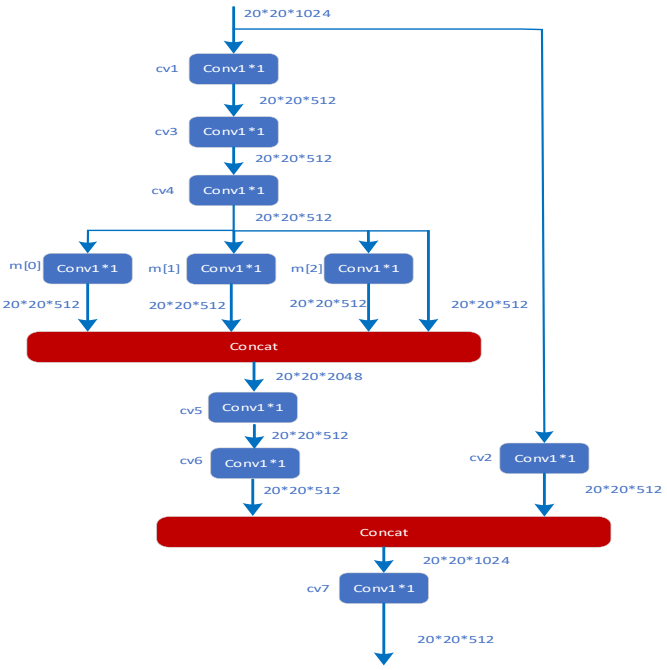
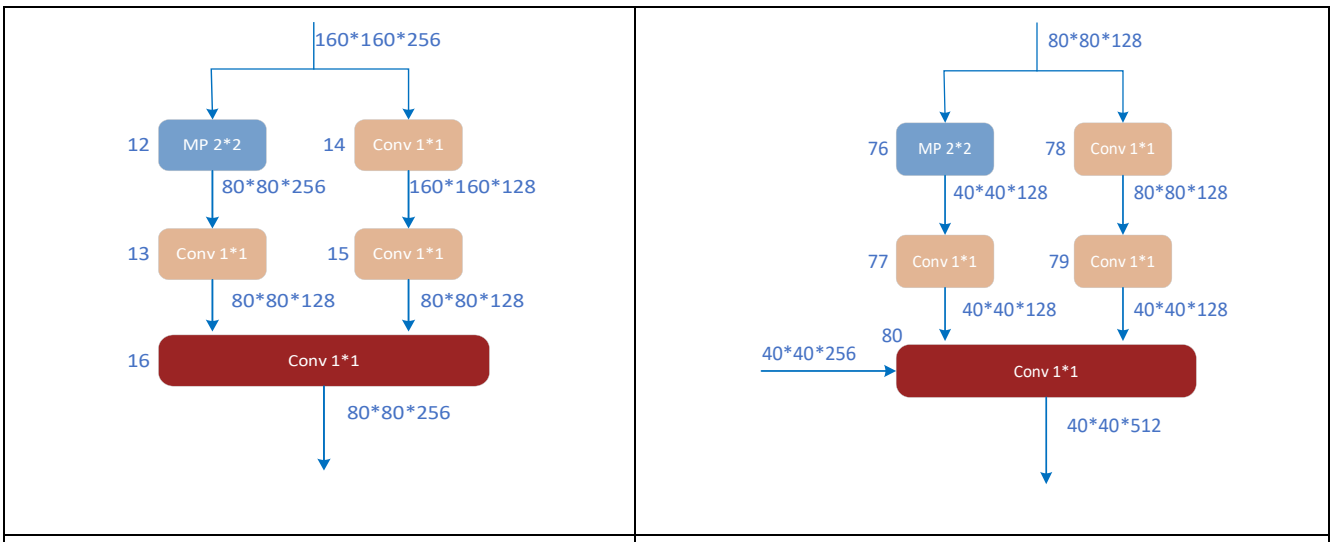


Figure 3. SPPCSPC



a MP-1 b MP-2

Figure 4. MP module

3. Experiments

In the experiment, ReCTs was used as the model training data set, with a total of 20000 images, of which 15000 were used as the training set and 5000 as the test set. After some preprocessing, the image is mapped into a 640 pixels high and 640 pixels wide image by affine transformation. Epoch is set to 300, and the initial learning rate is 0.01. The loss curve is shown in Figure 5. The final experimental map reached 78.36%.

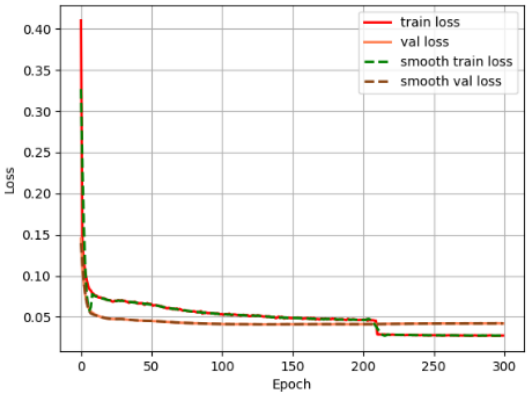


Figure 5. Model training loss

4. Conclusions

The text in images and videos contains rich and accurate high-level language descriptions. Accurate and effective extraction of these text information has important applications in multimedia retrieval, human-computer interaction, robot navigation, industrial automation and other fields. For the text location in Chinese scene, we use the target detection method to carry out end-to-end text location based on YOLOv7 model, take the street view image as input, detect the text line in the image area, and return the location information of the text line. The final map reached 78.36%.

Acknowledgment

Project supported by the Natural Science Foundation of Hebei Province (F2014502081) and the Fundamental Research Funds for the Central Universities(2020MS120)

References

[1] Judd T. Learning to predict where humans look IEEE international conference on computer vision[J]. Proc. ICCV, 2009, 2009.

[2] Breuel T M, Yanikoglu B A, Berkner K. The OCRopus open source OCR system (Proceedings Paper) [J]. IEEE, 2002.

[3] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005.

[4] Low D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004.

[5] Ojala T, Pietikainen M, Maenpaa T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[C]. IEEE, 2002: 971-987.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[7] Liu W, Berg A C, Fu C Y, et al. SSD: Single Shot MultiBox Detector[C]. ECCV, 2016: 21-37.

[8] Liao M, Shi B, Bai X, et al. TextBoxes: a fast text detector with a single deep neural network[C]. National Conference on Artificial Intelligence, 2017.

[9] Shi B, Bai X, Belongie S. Detecting oriented text in natural images by linking segments[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2550-2558.

[10] Ma J, Shao W, Ye H, et al. Arbitrary-Oriented Scene Text Detection via Rotation Proposals[J]. IEEE Transactions on Multimedia, 2017: 3111-3122.

[11] Wang C-Y, Bochkovskiy A, Liao H-Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.

[12] Ding X, Zhang X, Ma N, et al. RepVGG: Making VGG-style ConvNets Great Again[C]. IEEE, 2021: 13728-13737.