Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 1, 2023

Detection and Recognition Algorithm for Cigarette Box Documents based on Deep Learning

Qifan Wen

School of Qingdao University, Qingdao 266000, China

Abstract: In order to solve the problem of unclear detection and recognition caused by damage and distortion of the markings on the cigarette box during transportation, this article introduces an efficient channel attention mechanism SENet module for the DBNet detection model, effectively enhancing the model's detection ability. The introduction of deformable convolutional DCN on the CRNN recognition model further enhances the model's ability to detect distorted text. The experimental results show that the accuracy of the detection model reaches 88.43%, and the recall rate reaches 83. 47%, the comprehensive indicator reached 85.87%, with a significant improvement effect. The entire line recognition rate of the recognition model also reached 90.23%, which is better than the original algorithm.

Keywords: Text Detection; Text Recognition; Deformable Convolution; Deep Learning.

1. Introduction

With the continuous development of modern industrial production technology and information technology progress by leaps and bounds, mankind has entered the industry 4.0 era of the traditional mode of industrial production mode is also gradually to the transition to intelligent industrial production, innovation to create a "smart" factory has become the focus of the development of the field of industrial production direction.

The pace of iterative upgrading in the tobacco industry is also accelerating. In order to realize long-term, stable and healthy development, in the inspection line of cigarette case storage, in the past, the domestic production line generally adopts the method of manually inspecting the documents on the cigarette case, relying on the human eye to check the information marked on the documents of the cigarette case. The result shows that the correct rate of identifying the information of the cigarette case into the warehouse is greatly affected by human factors, and it is difficult to realize the complete warehousing and backup of the cigarette case in a stable manner, which also has a certain impact on the production efficiency of the enterprise. The traditional manual verification of cigarette box information in high-speed production, both labor-intensive, high misdiagnosis rate, far from being able to meet the needs of refined production. Currently, a cigarette factory tobacco library of the number of incoming and outgoing tobacco grades of various grades of tobacco, the daily number of incoming and outgoing boxes totaling about 1400 boxes, each box of cigarettes need to manually confirm the grade, year, material batch number, region and other information, while the cigarette box documents compared to most of the natural scenarios The data is more complex, mainly due to: 1, the paper is prone to smudges and wear and tear problems; 2, the documents template styles Complex and diverse; 3, text distribution is dense and there are many types of text; 4, stamped and misplaced resulting in text coverage overlap, interference is greater; 5, data collection is more difficult. These bring difficulties to the calibration of information. In particular, storage operators often need to work overtime to complete the day's task, and the lack of time series statistics and electronic information archiving, is not conducive to the measurement of the material beat and the establishment of the whole process of product traceability chain.

The purpose of this paper is to address the problems of the current algorithms in the detection and recognition task of cigarette box slips, and to explore model algorithms that can effectively solve the problem of detecting the text of cigarette box slips and improve the accuracy of the recognition of the text of the slips through the study of the text detection and recognition algorithms of the cigarette box slips in the natural scene by means of the deep learning method.

2. Detection Algorithms

2.1. DBNet Overview

The detection algorithm used in this paper is based on the DBNet [1] network and the differentiable binarization (DB) algorithm. The algorithm is trained to predict the threshold value of each pixel point by learning the threshold map and putting the threshold transformation into the network using a differentiable operation, instead of using a fixed value to separate the background from the foreground, thus solving the problem of non-differentiable gradient. The core idea of DBNet is to propose a differentiable binarization module block that incorporates the post-processing binarization process into the training to improve the segmentation effect and increase the inference speed at the same time. This improves the inference speed while improving the segmentation effect.

2.1.1. DBNet Model Structure

The structure of the DBNet model is shown in Figure 1. The input is an image, which is fed into a feature pyramid (FPN[2]) backbone network, then after upsampling the pyramid features are adjusted to the same scale, and cascade stitching (concat) to obtain the feature map Feature map, then the feature map is used to generate the probability map and threshold map, then after the joint operation of probability map and threshold map to obtain the approximate binary map of the network prediction, and finally using the bounding box generation module is used to convert the approximate The binary map is finally transformed into the form of candidate frame coordinates using the bounding box generation module.

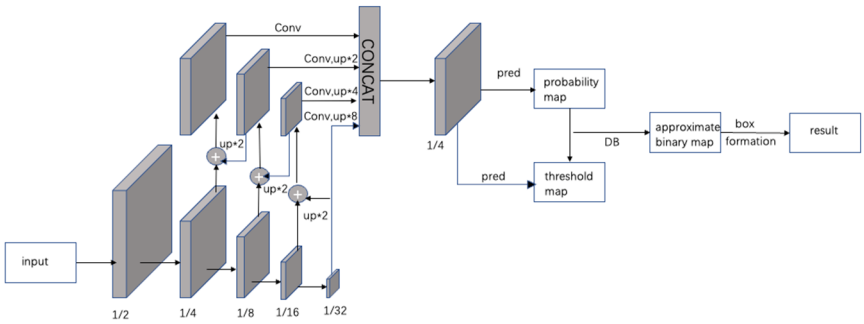


Fig 1. DBNet network model

2.1.2. Binarization

The segmentation network outputs a probability map $P \in R^{H \times W}$ of height H and width W. For ordinary binarization, the process is to set the pixel values in the positive sample regions to 1 and the values in the negative sample regions to 0. The computation procedure is shown in Equation (1).

$$B_{i,j} = \begin{cases} 1 & if \quad P_{i,i} \geq sth, \\ 0 & other \end{cases} \tag{1}$$

where *th* is the threshold for pixel classification. (i, j) are the position coordinates of the pixel points in the probability map. This binarization operation in the form of a step function is not microscopic and cannot be added to the network for gradient backpropagation during the training process. To address this problem, the standard binarization process can be replaced by an approximate step function as shown in (2).

$$B_{i,j} = \frac{1}{1+e^{-k\left(P_{i,j}-T_{i,j}\right)}} \tag{2}$$

where B denotes the approximate biplot, i,j denotes the truth value, $T_{i,j}$ is the adaptive threshold map for model prediction; k is the scaling weight.

2.1.3. Adaptive Thresholds

In DBNet, probabilistic and binary maps share supervisory information, while threshold maps require separate label generation. The labels are generated as follows: given an image with a detection instance, the original polygonal instance region is a collection of pixels, set to M, which needs to be shrunk to Ms. The offset of the shrinkage, D, is calculated by the perimeter, L, and area, A, of the original region, and the formula is shown in (3).

$$D = \frac{A(1-r^2)}{L} \tag{3}$$

Threshold map labels are generated by a similar process. This is done by first expanding the polygonal region M to Md with the same offset D as Ms. The gap between Ms and Md is regarded as the boundary of the instance region, which is represented in the threshold map as the boundary supervisory information of the instance.

2.1.4. Loss Function

The formula for the Loss function of DBNet is shown in (4):

$$L = L_s + \alpha \times L_b + \beta \times L_t \tag{4}$$

In this equation, L_s is the Loss of the probability map, L_b is the Loss of the binary map, and L_t is the Loss of the threshold map, and for α and β, the values are set to 1.0 and 10, respectively. for L_s and L_b can be solved by using the binary cross entropy (BCE) as follows:

$$L_s = L_b = \sum_{i \in S_i} y_i log x_i + (1-y_i)\log(1-x_i) \tag{5}$$

In this equation, S_i is the sampled dataset and the ratio of positive and negative samples is 1:3. L_t is computed in the following equation:

$$L_t = \sum_{i \in R_d} |y_i^* - x_i^*| \tag{6}$$

Where: R_d is all the pixels obtained by expanding the labeled box by the D offset, and y_i^* is the label of the computed threshold map.

2.1.5. Dbnet Improvements

In the process of feature extraction, in order to reduce the invalid features, we introduce the efficient channel attention mechanism SENet[3] module, and add the SENet module into the ResNet50 residual branch, which can make the network through the self-learning has been the importance and relevance of the different channels. The structure of SENet is shown in the following figure:

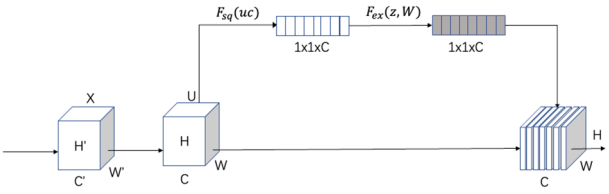


Fig 2. SENet Network Architecture

In the SENet module, the data of each channel is first pooled globally on average, in order to obtain the corresponding scalar, which is calculated as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_C(i,j), z \in R^C \tag{7}$$

Then after full connectivity, ReLU activation function, full connectivity, Sigmoid activation function the weight values distributed in the interval [0,1] and the channel importance are obtained. As in (8) Eq:

$$s = F_{ex}(z,W) = \sigma\big(g(z,W)\big) = \sigma(W_2\delta(W_1 z)) \tag{8}$$

The first fully connected layer is used for dimensionality reduction, r is the dimensionality reduction hyperparameter, and then activated using the ReLU [4] function, the next fully connected layer is used to recover the dimensionality, and then finally the s is obtained by the Sigmoid function, and finally, the raw data of each channel is multiplied by the weights computed from the above mentioned formulas, respectively, to obtain the weighted feature maps, as shown in (9):

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c . u_c \tag{9}$$

The text detection model combines SENet with ResNet50

by adding the SENet module to the residual learning branch of ResNet50 to form Res50-SENet.

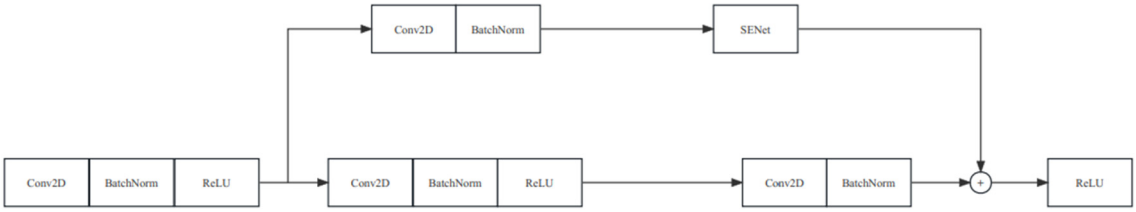


Fig 3. ResNet50 residual branch added to SENet

3. Recognition Algorithm

3.1. CRNN Network Architecture

CRNN[5] has a three-layer structure consisting of a convolutional layer, a recurrent layer and a transcription layer. Usually, CTC is acting in the transcription layer to achieve unaligned prediction of text. The overall network structure of CRNN can be seen from Figure 4:

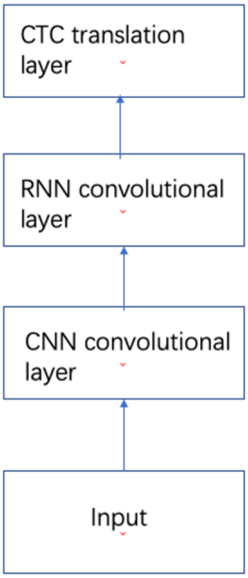


Fig 4. CRNN model

In this the convolutional layer uses a CNN network whose purpose is to extract the feature sequence. The bidirectional recurrent neural network (LSTM) used in the recurrent layer whose role is to predict the label distribution for each frame. And the transcription layer uses CTC to convert the label sequences into the final prediction.

3.1.1. Convolutional Layer

In the convolutional layer a deep convolutional neural network is used to extract the feature information of the input image as a way of obtaining the feature map, in which the input image is usually processed as a single-channel grayscale image, with the height set to 32 and the width to 160 When the size of the input image is 1 × 32 × 160, the output size of the convolutional layer is 512 × 1× 16. However, the feature map obtained at this time cannot be sent to the recurrent layer, it needs to be further converted (i.e., Map-to-Sequence), to transform the feature map into a set of feature vector sequences, which is done by combining each column of pixels into a vector according to the order of the generated feature map from left to right, and then arranging these vectors to obtain a set of feature vector sequences in sequence.

3.1.2. Recurrent Layer

The recurrent layer is the process of sequential inference on the features extracted from the previous convolutional layer. In CRNN, a Long Short Term Memory Network (LSTM) is used as the base network instead of the basic RNN, because RNNs can only consider the recent state and are not capable of preserving long term memories. In CRNN, the deep bi-directional LSTM[6] is used, which can be used to obtain long term information as well as higher level abstract information.

3.1.3. Transcriptome

The role of the transcription layer is to convert each vector in the probability matrix output from the loop layer into a labeled sequence, i.e., the labeled sequence with the highest probability is obtained. The characters are then aligned to eliminate the possibility that some characters may map more than one vector, etc.

3.1.4. Loss Function

The loss function used is the CTC[7] (Connectionist Temporal Classification) loss function, which is calculated as in (10):

$$L = \sum_{(I_i, l_i)\in\chi} \log(p(l_i|y_i)) \tag{10}$$

where χ=I_i, l_i denotes the dataset, I_i denotes the training image, $l_i$denotes the real labeled sequence, and $y_i$denotes the sequence of vectors outputted by the loop layer, where the sum of probabilities of all the vectors that can be mapped to I_i is calculated as shown in Equation (11)

$$p(l|y) = \sum_{\pi:\beta(\pi)=l} p(\pi|y) \tag{11}$$

Where π denotes the sequence of all vectors that can be mapped to l and β denotes the mapping relation.

3.2. CRNN Improvements

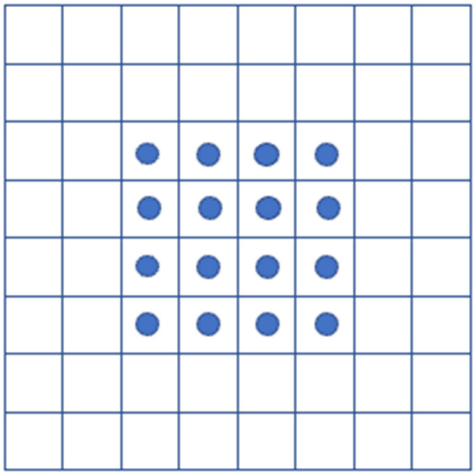


Fig 5. Standard convolution

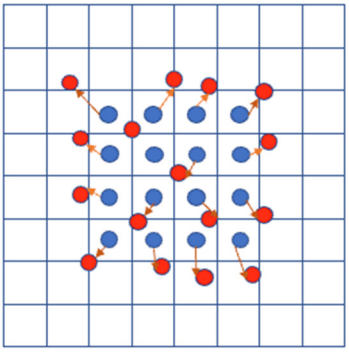


Fig 6. Deformable convolution

Although CRNN has achieved excellent recognition accuracy on regular image-text datasets, there is a problem of low recognition rate for some irregular images. This is due to the standard convolutional layer and then its sense of the field range of the sampling band your distribution is a kind of approximate uniform sampling, will be some of the text has nothing to do with the noise information is extracted, so we introduced to consider the introduction of deformable to extract the text features of the image.

Since deformable convolution [8] can change the geometry of the convolution kernel by adding learnable offsets to each sample point of the normal convolution kernel, as shown in Fig 5 and 6.

The operation of standard 2D convolution at each position p0 in the image can be expressed as:

$$y(P_0) = \sum_{P_n \in R} w(P_n) x(P_0 + P_n) \tag{12}$$

where x denotes the input image, w denotes the weights of the sampled points, and R denotes the fixed receptive field size and expansion coefficient of the convolution kernel. In deformable convolution, R can be expanded by a set of offsets $\{\Delta P_n | n = 1, \ldots, N\}$:

$$y(P_0) = \sum_{P_n \in R} w(P_n) x(P_0 + P_n + \Delta P_n) \tag{13}$$

The offset Δ of the sampling position can be automatically learned in the training stage, and the deformable convolutional feature sampling points can be better focused on the text area through learning, effectively reducing the extraction of irrelevant information such as background, thus effectively improving the recognition accuracy of irregular text images.

4. Experimental Analysis

4.1. Experimental Environment and Dataset

The text detection and recognition performed in this paper are trained on the same machine, the specific parameters used are CPU: Inter Core i7 9700 K, GPU: GeForce RTX 2080Ti, Memory: 16 GB, Operating System: Ubuntu20.04 64-bit, Programming Language: Python.

The text detection dataset is used is collected in a cigarette factory, a total of 1870 images were screened, using the label Img tool, the images are labeled according to the format of the VOC dataset, and divided into a training set, a test set, and a validation set, with the proportions of: 6:2:2, respectively.

The recognition dataset was generated by Text Recognition Data Generator on the basis of the detection dataset and a total of 34,202 images were generated.

4.2. Evaluation Indicators

In the detection algorithm, the evaluation metrics used in this paper are P (accuracy), R (recall) and F1 (composite metric). The formulas for different evaluation metrics are given below respectively:

$$P = \frac{TP}{TP+FP} \tag{14}$$

$$R = \frac{TP}{TP+FN} \tag{15}$$

$$F1 = 2\frac{P \times R}{R+P} \tag{16}$$

Where TP denotes the number of correctly predicted positive samples, FP denotes the number of incorrectly predicted positive samples, and FN denotes the number of incorrectly predicted negative samples.

In recognition algorithms, two criteria, the percentage of correct characters and the average edit distance, are usually used. However, for cigarette case documents, the error of a single character can lead to semantic errors, so it is necessary to ensure that the recognition of the text line is consistent with the real results. Therefore, we adopt the whole line recognition rate as the evaluation standard of text recognition, that is, each character in each line is recognized correctly before the text line is marked as correctly recognized, otherwise it represents a failure.

4.3. Detection Experiments

Table 1. The results of the experiments (1)

algorithms	iterations	P/%	R/%	F1/%
ResNet18	2000	81.03	74.34	77.54
ResNet50	2000	83.25	75.68	79.28
MobileNetV3	2000	75.65	72.12	73.36
Res50-SE-Net	2000	86.44	82.23	83.75
ResNet18	5000	82.57	74.87	78.53
ResNet50	5000	84.54	75.75	79.90
MobileNetV3	5000	80.13	73.33	76.57
Res50-SE-Net	5000	88.43	83.47	85.87

From the above experiments, it can be seen that when the number of iterations is 5000, the accuracy, recall and comprehensive indexes of the algorithm Res50-SE-Net proposed in this paper are 88.43%, 83.47% and 85.87% respectively, which are 5.86%, 8.6% and 7.34% higher than those of ResNet18, and 3.89%, 7.72% and 5.97% higher than those of ResNet 50. 3.89%, 7.72%, and 5.97%, and 8.30%, 10.14%, and 9.30% than the indicators of MobileNetV3, respectively. In summary, it can be seen that the Res50-SE-Net proposed in this paper outperforms the other algorithms in all the three indicators, which proves that the model proposed in this paper has better robustness.

4.4. Recognition Experiment

In the recognition task of cigarette box documents, since CRNN requires the height of the input image to be an integer multiple of 16, the image is uniformly set to 32, and then the model is allowed to use deformable convolution for the first convolution module after each downsampling, we experimented the text recognition model before and after the improvement on the dataset, and the results of the experiments are as follows:

Table 2. The results of the experiments (2)

algorithms	line recognition rate /%
CRNN	87.65
CRNN-DCN	90.23

In this paper, experiments are done on the benchmark model CRNN and the improved model CRNN-DCN, respectively. For CRNN the whole line recognition rate can reach 87.65%, while the recognition algorithm proposed in this paper, due to the incorporation of deformable convolution, the whole line recognition rate reaches 90.23%, which is 2.58% higher than the previous one, and the recognition accuracy is better improved.

5. Conclusion

In this paper, the following work has been accomplished mainly in text detection model and text recognition model:

(1) The cigarette box logo text detection model for the DBNet network model, the introduction of the SENet module, to further improve the feature extraction ability of the model, for the image of the font distortion, deformation and other fonts have a better ability to detect the improved model's accuracy, recall, the comprehensive indicators reached 88.43%, 83.47% and 85.87%, respectively, compared with the improvement of the previous model, there is a greater improvement. model, there is a big improvement.

(2) Cigarette box logo text recognition model for the CRNN network, the introduction of deformable convolution, due to the introduction of deformable convolution, for the broken text picture, there will be a better recognition effect, the accuracy of the improved CRNN model reached 90.23%, which is 2.58% higher than the CRNN model, indicating that this paper's improvement of the model is significant.

References

[1] Tok W H ,Stéphane Bressan. DBNet: A Service-Oriented Database Architecture[C]//17th International Workshop on Database and Expert Systems Applications (DEXA 2006), 4-8 September 2006, Krakow, Poland. IEEE Computer Society, 2006. DOI:10.1109/DEXA.2006.48.

[2] Schimanski L M , Drakesmith H , Merryweather-Clarke A T ,et al. In vitro functional analysis of human ferroportin (FPN) and hemochromatosis-associated FPN mutations[J].Blood, 2005, 105(10):4096-102.DOI:10.1182/blood-2004-11-4502.

[3] Hu J, Shen L , Sun G .Squeeze-and-Excitation Networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2018.DOI: 10.1109/ CVPR. 2018. 00745.

[4] Zhang C , Woodland P .Parameterised Sigmoid and ReLU Hidden Activation Functions for DNN Acoustic Modelling [C]// Interspeech 2015.2015.DOI:10.21437/Interspeech.2015-649.

[5] Mogren O .C-RNN-GAN: Continuous recurrent neural networks with adversarial training[J]. 2016.DOI: 10. 48550/ arXiv. 1611.09904.

[6] Sundermeyer M ,Ralf Schlüter, Ney H .LSTM Neural Networks for Language Modeling[C]// Interspeech. 2012. DOI: 10. 1016/0165-6074(89)90269-X.

[7] Graves A ,Santiago Fernández, Gomez F .Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[J].ACM, 2006.DOI: 10. 1145/ 1143844.1143891.

[8] Tomáš Jeníček. Deformable Convolutional Networks [J]. [2023-08-19].