Intelligent Bird's Nest Hazard Detection of Transmission Line based on RetinaNet Model

Zhang Hui1,*, Zhao Jian1, Chen Yuran1, Jiang Su1, Wang Di1 and Du Hao1

1Guizhou Electric Power Design&Research Institute, Power Construction Corporation of China, Guiyang 550000, China

* Corresponding author: 1554618332@qq.com

Abstract. Electric energy is a kind of secondary energy with clean and efficient, convenient and safe use and the most wide range of applications. The transmission line equipment and the bird's nest on the tower will adversely affect the transmission line equipment, and even endanger the reliable operation of the transmission line. At present, the traditional manual line patrol has the problems of heavy workload, high cost, low efficiency and many blind areas, and the classical machine learning algorithm for bird nest recognition and classification is relatively low in efficiency and accuracy. In order to solve this problem, the RetinaNet model based on deep convolution neural network is selected for automatic detection of bird's nest targets. By adjusting the appropriate network structure and parameters to optimize the model, a RetinaNet model suitable for bird nest detection is established. The experimental results show that the accuracy of RetinaNet is 94.1%, and the recognition speed of each image is 68ms. Compared with Faster R-CNN, YOLO and SSD methods, the validity and reliability of RetinaNet model for bird's nest detection on transmission line equipment and towers are verified.

1. Introduction

While birds improve the ecological environment and help human beings, they also bring trouble and harm to the power system. Nesting by birds will seriously pollute porcelain bottles, resulting in a decline in the insulation degree of insulators, resulting in insulator flashover and line short-circuit tripping [1]. In order to ensure the safe and stable operation of the transmission line as far as possible and minimize the harm caused by birds to the transmission line, it is necessary to detect the bird's nest. At present, the artificial interpretation of bird's nest images is inefficient, missed detection, and is often disturbed by subjective factors. The way of intelligent analysis and inspection of pictures has been the trend of the times.

At present, the algorithms used in intelligent target detection can be divided into two categories: the first is the second-order target detection model, which first extracts the candidate box, then extracts the image features through the convolution neural network (Convolutional Neural Network, CNN), and finally carries on the target classification and regression. The second type is the first-order target detection model, which directly transforms the classification and location of targets into regression problems without the need to extract candidate regions. The second-order target detection model can fully extract image features and achieve accurate classification and location. The accuracy of detection is higher than that of the first-order target detection model. Compared with the second-order target detection model, the first-order target detection model has simpler network structure, faster running speed and higher real-time performance. On the other hand, the real-time detection is required in the online inspection of the bird's nest, so it is more reasonable to use the first-order target detection model.
However, the dense detector of the first-order target detection model will encounter the extreme imbalance between positive and negative samples in the training process, which will affect the accuracy of bird's nest target detection. In this paper, combined with Retinanet algorithm to solve this problem successfully [2]. The experimental results show that the accuracy of this algorithm is significantly higher than that of other target detection algorithms.

2. RetinaNet model

The model network structure of Retinanet is shown in figure 1. The model consists of a feature extraction network ResNet, a backbone network and two task-specific subnetworks. The feature extraction network ResNet of figure 1 is responsible for extracting the preliminary features of the image; the function of the feature pyramid network FPN of figure 1 is to recombine the proposed preliminary image features with the high-level semantic feature information to complete the fine extraction of image features to enrich the receptive field of each feature layer; the sub-networks in figure 1 perform convolution object classification tasks and frame regression tasks on the output of the backbone network respectively.

![Figure 1: Schematic diagram of Retinanet network structure](image)

2.1 Feature extraction network

Deep residual network ResNet introduces the design of residual block. The principle of residual block is that the data output of the previous layers directly skips multiple layers and is introduced into the input part of the later data layer. It overcomes the problem that the learning rate becomes lower and the accuracy can not be effectively improved due to the deepening of the network depth. As shown in figure 2, it represents a residual block of ResNet.

![Figure 2: ResNet residual block](image)

As can be seen from the above figure, $X$ is the input value of the residual block. After passing through the first layer convolution layer and being activated by the activation function, the output $F(X)$, $F(X)$ is added to the input value of the residual block $F(X)$, after the activation function is activated before the $X$ is activated by the activation function after passing through the second layer convolution layer, so the residual block is also known as jumping or fast connection.

If a convolution layer is redundant, when there is no introduction of ResNet, each layer of learning needs to meet $H(X) = X$, which is difficult to implement. The structure of ResNet is
\[ F(X) = H(X) - X, \] and only \( F(X) = 0 \) is needed, which simplifies the problem, that is, the redundant layer can learn the update parameters of \( F(X) = 0 \) more quickly. Compared with ResNet, ResNet can transform learning goals into residuals through multiple jump connections and simplify learning goals and difficulties. ResNet structure includes ResNet-34, ResNet-50, ResNet-100 and so on [3].

2.2 Characteristic pyramid network
According to the characteristics of weak semantic information of low-level features, clear target location, strong semantic information of high-level features and fuzzy target location, FPN structure fuses the feature information of different layers through bottom-up connection, top-down connection and horizontal connection[4].

The structure of the main network part of RetinaNet is not completely consistent with that mentioned in FPN. The RetinaNet model only uses C3, C4 and C5 feature images to avoid generating Anchor, in high-resolution C2 feature images to reduce model detection time. The feature pyramid layer P6 is obtained by convolution of \( 3 \times 3 \) convolution kernel with step size 2 on the basis of C5, and P7 is obtained by adding a RELU to P6 and convolution with \( 3 \times 3 \) convolution kernel and step size 2.

2.3 Subnetwork and loss function
In RetinaNet, deeper convolution is used, with a total of five layers. In parallel with the classified subnet, the FPN output of each layer is connected with a location regression subnet, which is essentially a FCN network, which predicts the offset of anchor and its corresponding GT location. The first is also four layers of 256D convolution, and the last layer is the 4A dimension, that is, for each anchor, we return a four-dimensional vector of \( (x, y, w, h) \). Location regression has nothing to do with category. Although classification and regression subnets have a similar structure, the parameters are not shared.

In order to solve the problem of the imbalance between positive and negative samples caused by the candidate mechanism of dense sampling in the first-order target detection model, the Focal Loss function is used in this paper. This function is an improved cross-entropy loss function. By multiplying the original cross-entropy loss function by the exponential formula which weakens the contribution of the easily detected target to the model training, it successfully reduces the phenomenon that the target detection loss is easily controlled by a large number of negative samples. Its plan. The formula is as follows:

\[ f_{FL}(p_i) = -\alpha_t (1 - p_i)^\gamma \log p_i, \] (1)

In the formula: \( \alpha_t \) is the equilibrium factor, with a value of \([0, 1]\); \( p_i \) is the probability of belonging to the prospect predicted by the model, with a value of \([0,1]\); \( \gamma \) is an adjustment factor, with a value between \([0,5]\); the definition is as follows:

\[ p_i = \begin{cases} p & y = 1 \\ 1 - p & \text{Other} \end{cases} \] (2)

This loss function adds two factors \( \alpha_t \) and \( (1-p_i)^\gamma \) to the standard cross entropy standard. The imbalance between positive and negative samples can be suppressed by \( \alpha_t \), and the imbalance between simple and difficult samples can be controlled by \( \gamma \). Through the formula (1), whether it is the foreground class or the background class, the larger the \( p_i \) value is, the smaller the \( (1-p_i)^\gamma \) weight is. That is to say, simple samples can be suppressed by weight; \( \alpha_t \) is used to adjust the ratio of positive and negative samples, and when foreground categories use \( \alpha_t \), the corresponding background categories
use $1 - \alpha_i$. The optimal values of $\gamma$ and $\alpha_i$ influence each other, so it is necessary to combine them to get the optimal Focal Loss function when evaluating the accuracy.

3. Experiment analysis

3.1 Data set processing

The transmission line bird's nest data, including UAV patrol images, on-line monitoring equipment videos and on-site images taken by patrol personnel are collected, and a sample database is established. In the experiment, 2500 images are selected artificially, and 80% of them are used as training set and 20% as verification set according to the needs of the experiment. In order to prevent overfitting and increase the robustness of the whole network, each picture is flipped randomly, and the number of training samples is expanded to 5000. Select the complete and unobstructed, small deformation and good light pictures of the bird's nest that appear in the training sample. The data set generation mainly includes manually tagging all the images in the sample library by LabelImg software to generate XML tag files in accordance with the PASCAL VOC2007 format. The label file contains the following information about the picture: picture name, picture path, picture pixel height, width and depth, and the coordinates of the rectangular border of the area where the Bird's Nest is located in the picture $(xmin, ymin, xmax, ymax)$, is the upper-left and lower-right coordinates of the rectangular border. Finally, a script is written to convert the data in PASCAL VOC2007 format to csv data format to reduce the data storage capacity.

3.2 Experimental environment

Considering that the deep learning framework keras is easy to use and strong expansibility, the experimental part of this paper is based on this framework and accelerates the operation with the help of CUDA (Compute Unified Device Architecture). In addition, the experimental machine of this paper is a personal computer equipped with CPU of Intel (R) Core (TM) i7-8750H and GPU of NVIDIA GeForce GTX 2080. The size of memory and video memory are 32GB and 16GB respectively. The machine runs on Windows10.

3.3 Training process and evaluation methods

In order to avoid the uncertainty caused by the initialization of neural network weight, the residual neural network ResNet-50, which is pre-trained on the ImageNet data set, is taken as the backbone network of FPN. Some parameters are set as follows: the learning rate is 0.00001, the maximum number of training rounds is 50 rounds, and in FL, $\alpha_i = 0.25$, $\gamma = 2$.

Before the start of training, you need to set model training parameters, including picture width, picture height, the number of pictures used in each training step (batch-size), the total number of training rounds (epochs), and the number of training steps per epoch (steps). After setting, you can start training, and you will get multiple model weight files after the training is completed. The experiment was carried out on the RetinaNet model. The batch-size was set to 4, and 30 was set to 500.

The specific training steps and the training results of each step are as follows:

The main results are as follows:

1) the data set processed by 2.1 is inputted into the ResNet50 network structure to get the feature map of the original picture.

2) using the feature graph obtained in the previous step, the FPN network is constructed according to the method described in Section 1.3. the size decreases sequentially. The feature graph of any number of channels is selected on the FPN, and nine kinds of Anchor are generated for each sliding window through the RPN network, and then the candidate regions on the feature map of this layer are screened by non-maximum suppression.

3) input the Anchor into the classification sub-network and the frame prediction sub-network to train until the final network convergence, and complete the construction of the bird's nest detection model.
The weights and bias values of different convolution kernels in the network structure represent the feature results of the learning. The convergence curve of model training decreases with the increase of the number of iterations.

For the data set, the conventional indexes of accuracy and detection speed are used as the final evaluation index. In this experiment, on the basis of assuming that there are only positive and negative samples, the number of positive samples that are correctly classified is represented by $T_p$, the number of negative samples marked as positive samples by errors is represented by $F_p$, and $F_N$ represents the number of positive samples marked as negative samples by errors. Recall $P_{\text{recall}}$ and precision $P_{\text{precision}}$ are calculated according to formula (3) and formula (4).

\begin{equation}
P_{\text{recall}} = \frac{T_p}{T_p + F_N} \tag{3}
\end{equation}

\begin{equation}
P_{\text{precision}} = \frac{T_p}{T_p + F_p} \tag{4}
\end{equation}

Detection speed is another important index to measure the quality of target detection. Only with high speed can the problem of real-time detection be solved. Evaluate the speed of testing the time spent on each picture. Because different hardware will affect the detection speed of the picture, it needs to be compared and analyzed on the same hardware.

In this paper, the RetinaNet model is drawn with recall as Abscissa and precision as ordinate, and the area between curve and coordinate axis is the average accuracy. In this paper, even if the average accuracy and detection speed are used as the evaluation index of the model performance.

After several rounds of training, the average accuracy of the bird's nest has been greatly improved, but with the further increase of the number of training rounds, the detection model tends to be stable. Its value fluctuates around 0.9. When the number of samples of bird's nest category is small, it takes more rounds of training to stabilize the average accuracy. Therefore, the adverse effects of insufficient data samples can be alleviated by appropriately increasing the number of training rounds.

3.4 Experimental results and analysis

The bird's nest images with different angles and partial occlusion are selected for detection, and figure 3 is the final result of the detection. The specific accuracy of using this model to identify a bird's nest is shown in the following figure (the identified bird's nest is marked in a box, and the label above the box shows the probability of a bird's nest).

![](image)

a) Detection of bird's nest at the top of tower
b) Detection of bird's nest in the middle of tower

c) Bird's nest detection under top view

Fig3. Experimental result

It can be seen that this algorithm can effectively detect complex cases, effectively learn the features of data sets, and adaptively deal with difficult samples such as multi-angle, occlusion and so on. The main reasons are as follows: traditional methods such as directional gradient histogram mainly rely on artificially designed colors, gradients and other features, while neural networks can learn other features more comprehensively and have stronger representation ability for bird's nest targets. Secondly, compared with the second-order target detection model algorithm, the improved loss function can automatically increase the weight of difficult samples, reduce the role of easy-to-classify samples in the training process, and achieve a high-efficiency and difficult-to-train model. In order to compare the effect of RetinaNet detection, this paper uses the classical first-order target detection model and second-order target detection model. YOLO (You Only Look Once), SSD (Single Shot multibox Detector) and other methods are used in the first-order target detection model. Among them, the YOLO model has the defect of low detection accuracy [5], and the SSD method has shortcomings in small target detection [6]. The second-order target detection models include R-CNN (Region Convolutional Neural Network), Fast R-CNN (Fast Region Convolutional Neural Network), Faster R-CNN (Faster Region Convolutional Neural Network), in which R-CNN is based on convolution neural network (CNN), linear regression, and support vector machine (SVM) and other algorithms. CNN represents the image as a global high-dimensional vector, which has strong resolution in image search. R-CNN follows the traditional idea of target detection, only in the step of feature extraction, the traditional feature is replaced by the feature extracted by deep convolution network [7]. Fast R-CNN, under the CNN feature framework, uses fast regions to significantly improve the average accuracy of target detection [8]. The biggest difference between Faster R-CNN and Fast R-CNN is that a network called RPN (Region Proposal Networks) is
proposed, which is specially used to recommend candidate areas, which greatly improves the speed of
target detection [9]. Compared with R-CNN and Fast R-CNN, Faster R-CNN has obvious advantages
in detection accuracy and execution efficiency [10]. All the methods in the experiment run on the same
platform as above. The experimental results are shown in Table 1. As can be seen from figure 3, the
accuracy of RetinaNet model in detecting bird's nest is more than 90%, and some are even higher than
95%. It can be seen from Table 1 that the detection accuracy and speed of this RetinaNet model are
higher than that of the second-order target detection model Faster R-CNN. The detection speed of the
RetinaNet model is lower than that of the first-order detection models YOLO and SSD, but the detection
speed is lower than that of 100ms. When considering the detection speed and accuracy, the RetinaNet
model is an effective bird's nest target detection method.

Table1. Effect Comparison of different models on the data set of this paper

| Model name | Average accuracy (%) | Time required for each picture to be detected (s / piece) |
|------------|----------------------|--------------------------------------------------------|
| Faster R-CNN | 90.2 | 0.205 |
| YOLO | 88.5 | 0.034 |
| SSD | 89.3 | 0.045 |
| RetinaNet | 94.1 | 0.068 |

4. Conclusion

In this paper, the RetinaNet model based on ResNet residual network and loss function Focal Loss will
be used to detect the bird's nest on the tower and equipment. The appropriate network structure and
parameters are determined through a series of control experiments, and a RetinaNet model suitable for
bird nest detection is established. On the premise of ensuring the detection speed, the target detection
accuracy is improved, and the target detection accuracy is more than 90%. At the same time, through
the enhancement of the data set, the prediction accuracy of the model is improved. Compared with the
existing Faster R-CNN, YOLO and SSD methods, it is proved that the model has a good detection effect,
and can identify the bird's nest on the tower and equipment quickly and accurately, which is helpful to
improve the inspection efficiency and ensure the safe and stable operation of the transmission line. In
this paper, only 50-layer ResNet residual network is used for test and learning, and no more kinds of
attempts are carried out. Due to the fact that the network layer is not deep enough, its accuracy can be
further improved, and its generalization ability needs to be optimized by training model. At the same
time, the transmission line passes through all kinds of complex geographical environment, and the
network structure of the line is complex [11]. In the future, the existing data set will be expanded and
more in-depth experiments will be carried out.

References

[1] XU Jing, HAN Jun, TONG Zhigang, et al. Method for detecting bird’s nest on tower based on
UAV image[J]. Computer Engineering and Applications, 2017, 53(06):231-235.
[2] SUN Fuchun,KONG Tao,HUANG Wenbing,et al. Feature Pyramid Reconfiguration with
Consistent Loss for Object Detection.[J]. IEEE transactions on image processing : a
publication of the IEEE Signal Processing Society,2019,28(10):5041-5051.
[3] FANG Guokang,LI Jun, WANG Yaoru.Real-time face recognition on ARM platform based on
deep learning[J].Journal of Computer Applications, 2019, 39(08):2217-2222.
[4] ZHAO Tong,LIU Jieyu, SHEN Qiang.An Improved Multi-Gate Feature Pyramid
Network[J].Acta Optica Sinica,2019,39(08):235-244.
[5] WEN Jiewen, ZHAN Yinwei, LIN Weilin,et al. Batch re-normalization of real-time object
detection algorithm YOLO[J].Application Research of Computers,2018,35(10):3179-3185.
[6] TANG Cong, LING Yongshun,Zhengkedong,et al.Object detection method of multi-view SSD
based on deep learning[J].Infrared and Laser Engineering,2018,47(01):302-310.
[7] LIU Ruoyu, ZHAO Yao, WEI Shikui, et al. Indexing of CNN Features for Large Scale Image Search[J]. Pattern Recognition, 2015, 48(10):2983-2992.

[8] JIE Zequn, LIAN Xiaodan, FENG Jiasi, et al. Scale-aware Pixel-wise Object Proposal Networks[J]. IEEE Transactions on Image Processing, 2016, 25(10):4525-4539.

[9] REN Shaoqing, HE Kaiming, ROSS Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.

[10] SONG Huansheng, ZHANG Xiangqing, ZHENG Baofeng, et al. Vehicle detection based on deep learning in complex scene[J]. Application Research of Computers, 2018, 35(04):1270-1273.

[11] CHEN Liangqing, TANG Haicheng, XIAO Jianhua. Research on risk warning and recognition of transmission line based on deep learning[J]. Power Systems and Big Data, 2018, 21(12):1-5.