Object Detection in Optical Remote Sensing Images Based on Residual Network

Da Li, Shaoxing Gong and Dong Liu*
School of information engineering, Wuhan University of Technology, China

*Email: liudong@whut.edu.cn

Abstract. Conventional object detection algorithm based on deep learning only make use of deep feature which would become indiscriminative for small target in optical remote sensing images when the network deepens. In this paper, a multi-layer feature fusion method based on residual learning is proposed, which combine the shallow feature with deep feature to classify object comprehensively. First, a ResNet50 backbone network is constructed to extract feature from multiple layers. Then, the scales of feature at different layers are unified through RoI Pooling layer to screen region proposals and be classified through the SVM classifier. Comparative experiments are conducted on UCAS_AOD dataset released by UCAS and the result shows that our model achieved relatively good performance with 0.8802 and 0.9120 f1 score in car and airplane categories respectively.

1. Introduction
Object detection in optical remote sensing image is one of the research hotspots in the fields of computer vision, neural network and image processing. It is used to acquire accurate battlefield information in high-tech military confrontation, capture strategic targets, and provide accurate qualitative positioning information. Object detection is based on the extracted image feature, and related algorithm is used for object classification and location regression. Choosing appropriate feature extraction methods and matching classification strategies is the key to this process.

The construction of detection framework is closely related to feature extraction methods. At present, object detection models are mainly divided into two main categories, as follows:

(1) Traditional statistical-based feature extraction methods: In this kind of methods, the image is regarded as a random vector and target patterns are distinguished by statistical method. The typical methods are Histogram of Oriented Gradient [1] and Viola-Jones Object Detection Framework [2][3]. The purpose of these algorithms is to quickly calculate and predict feature on the premise of extracting rich and accurate feature. However, the feature extracted by traditional algorithms are basically low-level, manual feature, which are relatively more intuitive, easy to understand, more targeted to specific objects, but not generalizable.

(2) Feature extraction method based on neural network: Deep learning became the research focus rapidly after its excellent performance was discovered in image processing and a series of algorithms based on deep learning are proposed in object detection. In 2014, R-CNN [4] was proposed and achieved amazing effects on VOC07 dataset, and the mAP index has been increased from 33.7% to 58.5%. But its training and detecting speed are unsatisfactory. Later, SPPNet [5] proposed by Kaiming He et al. solved this problem effectively. SPPNet can increase the detection speed to 38 times of R-CNN without loss of accuracy. While SPPNet reduces the computing redundancy in region proposals,
it still have two main problems. First, its multi-stage training process is inefficient. Second, the network fine-tuning only takes place in the fully connected layer, which affects the accuracy of network. In order to further solve these problems, Girshick et al. proposed Fast R-CNN [6] characterized by a multi-task learning method realizing fine-tuning of the network and training classification and regression layers synchronously. Its training speed is 9 times and detection speed is 200 times comparing with R-CNN and it implements 70.0% of mAP on VOC07 dataset. Later, Faster R-CNN [7] proposed by Shaoqing Ren, Kaiming He, and Girshick et al. increases the detection accuracy to 78.8% due to its innovative design of region proposal network and multi-anchor mechanism. And the detection speed is also increased to near real-time speed.

Different functional modules such as feature extraction, region proposal and so on are gradually integrated into one network framework with the development of detection algorithm, but these are usually operated only using deep feature because it contains abundant semantic information. Meanwhile, the shallow feature which contains abundant location information is not utilized throughout the whole process.

In order to improve the accuracy of detection model, this paper proposes a feature fusion method based on residual learning. Feature maps extracted from multiple layers in ResNet50 are screened by region proposal network. Then, both the deep and shallow feature of each proposal are unified and fused to be taken advantage of in the classification and regression network to get the final detection results. Our method is tested in experiments conducted on the UCAS_AOD dataset [8], a remote sensing image dataset released by UCAS. The result shows that the model using feature fusion module achieved a good performance.

2. Related work
Related work associated with this paper will be introduced in this section including deep residual learning, region proposal network (RPN) and non-linear support vector machine classifier (SVM).

2.1. Residual learning
In ILSVRC 2015, a neural network with 152 layers trained by Kaiming He et al. took the 1st place achieving a top 5 error rate of 3.57% [9] which is an excellent performance. This network is built on plenty of residual units in the theory of residual learning which can not only accelerate the ultra-deep network training process but also greatly improve the accuracy of detection model.

Assuming that \( x \) is the input of a certain neural network and \( F(x) \) is the expected output. If the initial result of output is set to \( x \) by a shortcut, then the residual result \( R(x) \) that network still need to learn develop into \( R(x) = F(x) - x \), as shown in figure 1, a residual unit. ResNet equivalently changes the learning goal from the original output \( F(x) \) to the residual part between input and output \( F(x) - x \). ResNet solves some problems exist in conventional CNN such as information loss and deterioration in information transmission. The integrity of the input information is protected through a shortcut therefore the learning goal degenerates into the residual part which simplifies the learning objective and difficulty.

2.2. Region proposal network
Regional proposal network (RPN) is designed to generate region proposals in object detection framework, and it is firstly introduced in Faster R-CNN. Before RPN came into being, the Selective Search algorithm was usually used to generate candidate boxes in object detection frameworks, which was a traditional method and time-consuming. It took 2 seconds to complete a graph in CPU. The advantage of using RPN is that on the one hand, RPN takes less time, on the other hand, RPN can be easily integrated into Fast R-CNN and become a whole framework. The introduction of RPN can really integrate the whole process of object detection into a neural network.
Figure 1. Residual learning: a residual unit.

Figure 2. Region proposal network.

Figure 2 shows the process of RPN. After sliding window processing, a fully connected feature map is generated with dimension 256 (ZF network) or 512 (VGG network), which is input into two subsequent fully connected layers (reg layer and cls layer). Note that all spatial locations share the fully connected layer weights since the network works as a sliding window. This network is composed by an n*n convolution layer followed by two 1*1 convolution layers.

**Anchors:** At each position of sliding window, multiple region proposals are calculated simultaneously. Assuming that at each position, the RPN output k predictions, then the classification layer output 2k scores to estimate the probability that each region proposal is the target or the background and the regression layer output 4k coordinates to encode the location prediction of proposals.

**Loss Function:** To train RPNs, every proposal has a label to indicate that it is a target (positive) or background (negative). The two kind of proposals will be labeled as positive: 1) the highest IoU ratio with the truth box. 2) The IoU ratio with a certain truth box is not less than 0.7. Therefore, multiple positive labelled proposals can correspond to one truth box. A proposal would be set to negative if it has no IoU ratio greater than 0.3 to all truth boxes. These samples between positive and negative samples do not contribute to training. The following equation is the objective function that need to be minimized in the training process:

$$L(\{p_i\}, \{t_i\}) = \xi \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) + \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*)$$

(1)

i is the ordinal number of anchors in each batch in the training, $p_i$ is the probability of anchors.
being target, \( p_i^* \) is the sample label (only 0 or 1), \( t_i \) are the four parameters of the prediction box and \( t_i^* \) are the parameters of the labelled box. The coordinate positions of the region proposals satisfy the following relationships:

\[
\begin{align*}
    t_x &= (x - x_a) / w_a, t_x^* = (x^* - x_a) / w_a \\
    t_y &= (y - y_a) / h_a, t_y^* = (y^* - y_a) / h_a \\
    t_w &= \log(w / w_a), t_w^* = \log(w^* / w_a) \\
    t_h &= \log(h / h_a), t_h^* = \log(h^* / h_a)
\end{align*}
\] (2)

\( L_{cls} \) is classification loss function and \( L_{reg} \) is regression loss function, \( p_i^* L_{reg} \) means that regression loss only count positive samples (for negative samples \( p_i = 0 \). Classification layer and regression layer output \( p_i \) and \( t_i \) respectively. These two parts are normalized by \( N_{cls} \) (batch size is 256) and \( N_{reg} \) (anchor location is 2400) and weighted by a balance parameter \( \xi \). By default, \( \xi = 10 \), so the loss weights of classification and regression are roughly the same.

2.3. Nonlinear Support Vector Machine

Support vector machine(SVM) is a classical classification model which achieved excellent performance in binary classification especially. For the best classification accuracy, the hyperplane constructed to separate samples must have the biggest margin with the nearest sample.

For a given sample \( T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\} \) where \( x_i \in \mathbb{R}^n \), \( y_i \in \{-1, +1\} \), \( i = 1, 2, 3, ..., N \), if the sample points cannot be completely divided into two parts by a linear hyperplane, then a nonlinear SVM works by mapping the original input to a high-dimensional feature space through a kernel trick.

If there is a linear classification function satisfying requirements, the optimal hyperplane can be constructed in this high-dimensional space. The classification function can be obtained by following steps:

1. Choosing the appropriate kernel function \( K(x, z) \) and parameters \( C \) to construct and solve the optimization problem:

\[
\min_\alpha \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N \alpha_i
\] (3)

Subject to:

\[
\sum_{i=1}^N \alpha_i y_i = 0 \] (4)

\[
0 \leq \alpha_i \leq C, i = 1, 2, ..., N \] (5)

The optimal solution \( \alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)^T \) can be obtained by using quadratic programming;

2. We can find some \( i \) that satisfy \( 0 < \alpha_i^* < C \) and then solve

\[
b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i K(x_i, x_j)
\] (6)

3. Finally

\[
f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x_i, x) + b^*\right)
\] (7)
When $K(x,z)$ is positive definite kernel function, the problem (3) (4) (5) is a convex quadratic programming problem, and the solution exists.

3. Object Detection Based on Fused Feature

This section mainly introduces the basic principle of feature fusion method and the process of remote sensing image object detection based on fused future.

3.1. Comprehensive Classification Based on Fused Feature

Conventional object detection models (for example, Faster R-CNN) usually extract feature maps through a backbone network, and then send them to RPN network and the following classification and regression layers. However, after several max pooling operations in backbone networks, the feature map will become indiscriminative. For example, a 32*32-pixel object, after 4 maximum pooling operations in the network, the final feature map is only 2*2-pixel size.

Considering that the shallow feature of convolutional neural network containing abundant spatial information and the deep feature contain abundant semantic information, the accuracy of the model will be improved if these multi-level feature are combined for classification. Therefore, a feature fusion method is proposed in this paper, which extracts the output feature of the third, fourth and fifth convolution layers from the network, pools them separately, unifies the feature size, and then combines the feature of each layer into one output through a concatenate layer and sends them to SVM classification.

![Figure 3. The overview of our algorithm.](image)

3.2. The Execution Process of the Algorithm

Our detection framework is executed as the following steps:

1. Pre-process the picture and unify the size of the picture to 512*512*3;
(2) Input the pictures into the backbone network Resnet50 and get the output feature of the third, fourth and fifth convolution layers (called F3, F4 and F5);

(3) Input F5 feature into RPN network. At each sliding window position (after four times max pooling, F5 has 32*32 sliding window positions), k candidate boxes are predicted which meaning that the classifier layer outputs 32*32*2k classification prediction and reg layer outputs 32*32*4k coordinate regression prediction;

(4) Non-maximum suppression (NMS) algorithm is applied to the above prediction results, and 150 candidate boxes with the highest score and the lowest overlap (Intersection over Union, IoU) below 0.2 are retained;

(5) According to the regression positions of 150 candidate boxes, their corresponding F3, F4 and F5 feature are obtained. The feature sizes of each layer are unified by RoI Pooling layer, and the feature of each layer is connected by a concentrate layer;

(6) The fused feature is fed into SVM classifier and box_reg regression layer, and the final classification results and border regression values are obtained.

The flow of object detection is shown in figure 3.

4. Experiment and analysis
The experiment is carried out on the UCAS_AOD dataset performing object detection in car and airplane, both of which are very important targets in the military field. The dataset is annotated by UCAS. The car category contains 7115 targets in 510 pictures, and the aircraft category contains 7482 targets in 1000 pictures. The hardware platform includes i5-8400 CPU, 16GB memory and 1080Ti 11GB graphics card. Keras based on TensorFlow is installed to build the experimental environment in PyCharm.

F1 score is introduced to measure the performance of algorithms. As a statistical index, f1 score takes both precision and recall rate into account comprehensively which are frequently used to measure the performance of a model in binary classification problems. It can be calculated in the following ways:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$  (8)
The UCAS_AOD dataset is stochastically divided into three parts, 80% training set, 10% validation set and 10% test set. Figure 4 shows the curve of total loss in validation set and f1 score in test set with the increase of training epochs (1000 pictures per epoch).

Table 1 shows the performance comparison of several algorithms in UCAS_AOD dataset. Our method can reach 0.8802 in f1 score of car category and 0.9120 in airplane category, relatively 1%-2% higher than other advanced algorithms in performance.

Table 1. Performance comparison of algorithms on UCAS_AOD.

| Methods               | f1@0.2-car | f1@0.2-airplane |
|-----------------------|------------|-----------------|
| Our method            | 0.8802     | 0.9120          |
| Faster R-CNN-Resnet50 | 0.8697     | 0.9018          |
| Fast R-CNN            | 0.7733     | 0.8093          |
| YOLOv2                | 0.8595     | 0.8995          |
| SSD                   | 0.8442     | 0.8835          |
| RetinaNet             | 0.8617     | 0.8973          |

5. Conclusion

In order to aim at the small objects in optical remote sensing image, a feature fusion method is proposed in this paper. The shallow feature containing a lot of spatial information are combined with deep feature in convolution network for classification, solving the inadaptability of conventional detection model in small object detection. A comparative experiment carried out on UCAS_AOD dataset proves that our method improves the performance of detection model by 1% - 2%.

A feature fusion module is added to the model, which would increase the input parameters of the classifier and the training time of the network. How to simplify the network model and optimize the input parameters to reduce the training time of the model will be the focus of our future work.

References

[1] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2005.
[2] Viola P, Jones M. Rapid Object Detection Using a Boosted Cascade of Simple Feature [C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2001,1:511.
[3] Viola P, Jones M J. Robust Real-Time Face Detection [J]. IEEE International Conference on Computer Vision(ICCV), 2004:747-747.
[4] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]. Computer Vision and Pattern Recognition. IEEE, 2013:580-587.
[5] K. He, X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in European Conference on Computer Vision (ECCV), 2014.
[6] R. Girshick, Fast R-CNN, in IEEE International Conference on Computer Vision (ICCV), 2015.
[7] 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, 2015, 39(6):1137-1149.
[8] Zhu H, Chen X, Dai W, et al. Orientation robust object detection in aerial images using deep convolutional neural network[C]. 2015 IEEE International Conference on Image Processing (ICIP).
[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2, 4, 5, 6, 8.