Detection method of dense bridge disease targets based on SE-YOLOv3

Yaoling Wang¹, Zhi Zhang²* and Hang Yin³

¹ College of Computer Science & Technology, Wuhan University of Science & Technology, Wuhan, 430065, China
² College of Computer Science & Technology, Wuhan University of Science & Technology, Wuhan, 430065, China
*Corresponding author’s e-mail: zhangzhi@wust.edu.cn

Abstract. In this paper, we find that the overlapped detection frames of dense disease in the bridge disease data set, combined with the current research on target detection algorithms in dense scene. Based on YOLOv3, the feature pyramid is optimized to manually set thresholds for post-processing after extracting features. A bridge disease target detection model based on SE-YOLOv3 was proposed. In the post-processing process, IoU uses Soft-IoU to detect the accuracy of the prediction box, approves the value of IoU, and then uses the EM algorithm based on the Gaussian distribution to perform a function operation to delete the low-score prediction detection box to achieve the purpose of removing overlap. Experimental results show that the detection accuracy is 95.73%.

1. Introduction
The YOLOv3 algorithm performs well on a self-made bridge disease data set through effective joint feature pyramid multi-fusion training and new anchor points of k-means clustering, plus multi-scale training of the last full convolution. The fusion of three-level feature layers allows disease to be accurately labeled and classified accurately for both large and small targets. The residual network introduced in Darknet53 optimizes redundant regression pyramids, controls the propagation of gradients, and prevents gradients from disappearing or exploding. It is not conducive to training; the logistic regression classifier improves the detection speed based on the calculations made by suppressing the excessive number of background classes. Therefore, its detection efficiency is guaranteed.

However, the bridge disease still has the problem of densely stacked diseases, that is, the bounding boxes of similar objects that are very close to each other in a dense target scene are not accurately predicted and easily overlap. In addition, there are a lot of such environments in the dense target scene in the real world, but they are rarely included in the current neural network training / detection data sets. As a result, the performance of existing models tends to drop significantly on such datasets. At present, many of the datasets used in deep learning are still far from what the product is really facing - most of the pictures of object recognition are based on the target identifier, and the picture content is clean, even if the picture is based on the real world scene, they show is also a picture with a clear background.
Based on the above analysis, this experiment refers to Eran Goldman's idea of adding FPN to the Soft-IoU layer, combined with mixed Gaussian distribution units, using Expectation Maximization (EM) to solve overlapping detections and merge them into a single detection. This chapter builds a SE-YOLOv3 model based on bridge disease dense target detection. Soft-IoU is newly added to the backbone neural network, and then the EM algorithm (EM-Merger) is used to convert the Soft-IoU into a Gaussian distribution during the prediction process. The purpose is to merge multiple overlapping boxes to improve detection accuracy.

2. SE-YOLOv3 model design

2.1. Post-process

When we introduce non-maximum suppression, we first introduce a concept: Union-over-Union, IoU. It is the overlap ratio between the candidate frame and the original labeled frame generated in the target detection, so it is called the cross-ratio. The larger the value of IoU, the higher the overlap rate between the candidate frame and the candidate frame, which also means that the detection is more accurate, and the ideal value is 1. However, many candidate frames are generated during target detection training. Normally, a threshold is set to delete certain candidate frames. This is the NMS.

The NMS algorithm sets a parameter, assuming k. The candidate frames are classified according to the classification probability of the classifier. By starting with the frame with the highest probability (that is, the frame with the largest area), let M determine whether the IoU of other frames is greater than k. Then find the box with the second highest probability and repeat the process. This calculation method completely depends on the artificially set k value. If it is too large, too many probability boxes will be deleted and missed detection will be caused. Too small will not achieve the purpose of reducing overlapping boxes. As a result, extensive thinking has occurred in the field of object detection. This section lists more representative ideas for improvement:

1. Soft-NMS: Its core idea is to reduce confidence. After the IoU reaches the threshold, it is not deleted directly, but the confidence is reduced to enter a new iteration. Confidence is reduced using a function calculation, but the new threshold is still manually specified. The author mentioned that for SDD, the improvement of YOLOv2 is not large.

2. IoU-Net: Based on Soft-NMS, a RoIooling is proposed. The author mainly aims at the RoIs generated by RPN in Faster-RCNN. Randomly perturbate to get Jittered RoIs, filter out the boxes with GT's IoU <0.5, and use the rest as input to give IoU predictor via PrRoI Pooling. At the same time, IoU-guided NMS is proposed, which combines IoU with classification confidence. The final accuracy improvement is about 2%. But insufficient generalization ability

Synthesizing the above two viewpoints and reading a lot of literature, this article draws on the design of Soft-IoU. Its good performance on Retinanet and a more streamlined algorithm implementation are its highlights. By effectively improving the NMS process, the algorithm in this paper can be taken to a higher level.

2.2. SE-YOLOv3 overall architecture

Figure 1 shows the overall architecture of the SE-YOLOv3 model in this chapter. After inputting the image, feature extraction is performed through the backbone layer. Then, three FPNs are used to fuse multi-scale features to improve feature extraction performance. One is detection to generate a quadruple bounding box for each object, (x, y, h, w). They are the center coordinates of the bounding box, the height and width of the bounding box; the other is classification, which is used to predict the label C that indicates the presence or absence of a detection target; the last one is the Soft-IoU proposed in this paper.

With the addition of Soft-IoU, re-checking and evaluation is performed on the basis of the original IoU calculation. The purpose is to prevent IoU's greedy calculation method from filtering too many overlapping boxes. Confidence thresholds are sorted by confidence after IoU calculation, and
overlapping boxes are deleted according to the thresholds to achieve the effect of effectively eliminating overlaps. The calculation method of Soft-IoU will be explained in detail below.

![SE-YOLOv3 overall architecture](image)

Figure 1. SE-YOLOv3 overall architecture

### 3. SE-YOLOv3 model implementation

#### 3.1. Soft-IoU Layer

In a general object detection algorithm, the detected frame is subjected to NMS (Non-Maximum Suppression) post-processing. This post-processing is calculated based on the category, score, and IoU of the predicted frame. The score of the prediction box is generally used to predict the objects in the box. It does not represent the IoU between the prediction box and the real box, so there may be a high prediction box score, but the prediction box does not locate the object well. So, Soft-IoU is to estimate whether the positioning of the box is accurate, and whether the output of this layer is the IoU value of the predicted box and the real box (see Equation (3-1)).

\[
IoU_i = \frac{\text{Intersection}(\hat{b}_i, b_i)}{\text{Union}(\hat{b}_i, b_i)}
\]

Where \( \hat{b}_i \) is the predicted bounding box, \( i \in \{1..N\} \). \( b_i \) is the actual bounding box.

#### 3.2. Gaussian distribution

For the output of the network, there are N prediction frame positions, confidence levels, and predicted IoU scores \( c_{iou} \). In order to handle dense scenes, EM-Merger functions like NMS, and filters low-quality prediction frames in a similar way to k-means clustering. It has several steps:

1) The bbox output from the network is converted into a two-dimensional Gaussian distribution, that is, the Gaussian distribution in the horizontal and vertical dimensions of the image. One box corresponds to one Gaussian distribution. For N prediction boxes, use the following expression:

\[
F = \{f_i\}_{i=1}^{N} = \{N(p; \mu_i, \Sigma_i)\}_{i=1}^{N}
\]

2) where \( p \in \mathbb{R}^2 \) is two-dimensional. \( \mu_i = (x_i, y_i) \) means representing the gaussian distribution are represented using the central point of the prediction box. \( \Sigma_i = [(h_i / 4)^2, 0, 0, (w_i / 4)^2] \), \( (h_i, w_i) \) represents the length and width of the prediction box. For all prediction boxes, Mixed Gaussian (MOG) can be used as follows:

\[
f(p) = \sum_{i=1}^{N} \alpha_i f_i(p)
\]

(3-3)
Among them \( f_i \in F \), \( \alpha_i = \frac{C_i^{iou}}{\sum_{k=1}^{K} C_k^{iou}} \)

3) Because many bbox are overlapped in dense scenarios, resulting in overlapping of transformed gaussian distributions in the first step. So consider the mixed Gaussian distribution composed of less Gaussian distribution to represent the original mixed Gaussian distribution. That is, the mixed distribution composed of \( N \) original bbox transformed Gaussian distribution is replaced by the mixed distribution composed of \( K \) (\( K<<N \)) new Gaussian distribution.

\[
G = \{ g_j \}_{j=1}^{K} = \{ N(p; \mu'_j, \Sigma'_j) \}_{j=1}^{K}
\]

where the mixed distribution composed of \( K \) gaussian distributions needs to be similar to of \( N \) Gaussian distributions. \( K \) Gaussian distribution consists of the following:

\[
g(p) = \sum_{j=1}^{K} \beta_j g_j(p)
\]

To let \( f(p) \sim g(p) \), use KL-divergence solution, the formula is as follows:

\[
d(f, g) = \sum_{i=1}^{N} \alpha_i \min_{j=1}^{K} KL(f_i \parallel g_i)
\]

3.3. EM algorithm

Based on the similarity problem proposed in 3.2, referring to the idea of k-means clustering algorithm, a EM algorithm is proposed to solve \( K \) gaussian distribution by constantly excluding the nearest pair of gaussian distribution tuples until the remaining \( K \) are taken as the initial cluster center.

EM algorithm is divided into two steps, E-step and M-step. where E-step assign each bounding box to the nearest bounding box cluster, where the bounding box similarity is defined by the KL between the corresponding gaussian distributions, as follows:

\[
\pi(i) = \arg \min_{j}^{K} KL(f_i \parallel g_i)
\]

next, M-step will adjust the parameters for the new cluster samples, the expressions are as follows:

\[
\beta_j = \sum_{i \in \pi^{-1}(j)} \alpha_i
\]

\[
\mu'_j = \frac{1}{\beta_j} \sum_{i \in \pi^{-1}(j)} \alpha_i \mu_i
\]

\[
\Sigma'_j = \frac{1}{\beta_j} \sum_{i \in \pi^{-1}(j)} \alpha_i (\Sigma_i + (\mu_i - \mu'_j)(\mu_i - \mu'_j)^T)
\]

as the above procedure iterates until \( d(f, g) < 10^{-10} \) in formula (3-6), it is considered that the iteration converges. So \( K \) estimated Gaussian distribution is obtained, and the \( K \) range is:

\[
K = \text{size}(I) / (\mu_w, \mu_h)
\]

the \( K \) gaussian distributions after convergence are turned back bbox, the final prediction results are obtained. This process is unsupervised learning does not participate in training, so it is not calculated in the model training time.

3.4. Model training
The SE-YOLOv3 model training in this chapter uses the deep learning server of the Research and Development Department of the Bridge Science Research Institute. Remote training is implemented through Teamviewer. The experimental hardware environment configuration is shown in Table 1.

Table 1. Experimental Hardware Configuration

| Name          | Model                  | Num |
|---------------|------------------------|-----|
| CPU           | Intel XeonE5-2600      | 2   |
| slug          | Intel C602             | 1   |
| internal storage | 32GB DDR4 RECC      | 1   |
| video card    | NVIDIA RTX 1080 TI     | 2   |
| SSD           | 250GB SSD              | 1   |

The EM algorithm in training has a fast calculation rate in the 2D space matrix. The axis alignment of the detection frame improves the calculation speed of the diagonal covariance matrix. The positive IoU is set to greater than 0.5, and the negative IoU is set to less than 0.4. The score threshold of the detection frame was set to 0.05. After training for 40,200 iterations, the avg loss converged to 0.1286.

4. Experimental setup and result analysis

4.1. Experimental data and evaluation criteria

The data set used in the experiment is a self-made bridge disease data set. LabelImg image annotation software is used to annotate the picture frame and generate a .voc format training file. From this, some intensive diseases were selected as the key data sets of this experiment.
Figure 3. Bridge Disease Data Set

Figure 4. Labellmg

The evaluation of the experiment is a standard COCO Keypoint estimation metric as an evaluation index of the detection effect, plus Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), used to judge the model. The smaller the numerical value, the better the fitting effect, and the better the model's comprehensive ability. See (4-1) for both formulas:

\[
MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \\
RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}
\]  

(4-1)

4.2. Analysis of results

This article selects the current target detection accuracy and speed in the leading position of the algorithm to compare experiments to detect the improvement of the model
(1) Faster R-CNN: Use a small area proposal network RPN to reduce the number of anchor boxes and improve the speed of image processing. The region proposal network consists of an extended convolution layer (sliding window), a bounding box classification layer (box-classification layer), and a bounding box regression layer (box-regression layer). Among them, the extended convolution layer is used to reduce the dimensionality of the original feature layer, the bounding box classification layer is used to distinguish the object from the background, and the bounding box regression layer is used to roughly adjust the anchor box. This is the key to Faster R-CNN's high accuracy.

(2) Retinanet: Aiming at the problem of category imbalance of one-stage algorithm, a new loss function called focal loss is proposed, integrating the FPN and RPN ideas, and deleting the P2 layer and adding the P6 and P7 layers based on the FPN. It reduces the amount of calculation and improves the detection accuracy of large targets.

(3) YOLOv3: unique darknet53 network structure, combined with k-means clustering and logistic classifier to implement multi-classification and real-time accurate positioning, multi-scale fusion to achieve small target detection.

(4) SE-YOLOv3: This article proposes an algorithm that combines soft-IoU and EM to optimize the post-processing process.

Table 2. Comparison of models in test sets

| Method      | mAP  | MAE    | RMSE   | FPS  | Recall | F1-score |
|-------------|------|--------|--------|------|--------|----------|
| Faster-RCNN | 94.72| 107.46 | 113.42 | 2.37 | 0.9631 | 0.9176   |
| Retinanet   | 92.56| 26.584 | 43.962 | 15   | 0.8847 | 0.8754   |
| YOLOv3      | 90.84| 87.187 | 96.309 | 22   | 0.8963 | 0.9067   |
| SE-YOLOv3   | 95.73| 15.652 | 30.992 | 21   | 0.9375 | 0.9439   |

We can see that compared with several other text detection algorithms, the SE-YOLOv3 performs best in accuracy, and the comprehensive F1-score reaches 94.39%, in which the recall rate and accuracy rate are improved compared with the algorithm in Chapter 3, and the comprehensive F1-score is improved by 1.03%. Contrast detection speed, because there is no change to cluster location anchor and classification, so the SE-YOLOv3 speed is not different from the previous algorithm. On the regression, it can be seen that after adding Soft-IoU and EM, it decreases obviously in MAE and RMSE, and the model as a whole the fitting effect is better. Experiments show that SE-YOLOv3 can show more accurate detection on bridge disease data sets and also perform well in dense disease scenarios, which has good application value.

Compared with several other object detection algorithms, the SE-YOLOv3 advantage is to reduce the overlap of detection boxes:

Faster-RCNN as the two-stage algorithm, the RPN network and the Fast R-CNN network are trained separately, and the number of candidate boxes is greatly reduced by two classification, anchor the redundancy of convolution calculation is reduced, and then integrated on the Fast R-CNN. Because of the lack of utilization of the middle feature layer and the deepening of the network structure to obtain higher accuracy, the detection speed is twice as slow as the one-stage, and it takes hundreds of milliseconds to detect a picture, which is not suitable for real-time scenes. SE-YOLOv3 algorithm based on YOLOv3 Yes one-stage, direct prediction is used to keep the detection speed fast when combining Soft-IoU and EM.

Retinanet algorithms, like YOLOv3, are one-stage based algorithms, which are based on the structure of the Resnet50 network. The addition of focal loss new loss function solves the problem of unbalanced one-stage algorithm categories, makes the number of positive and negative samples tend to be balanced, the error function converges, and the addition of FPN improves the accuracy of feature extraction. The Retinanet algorithm, however, is much less rapid than YOLO. because it increases more anchor box, in raising the recall. Therefore Even if the accuracy difference is small, the speed is too slow model overall effect is not good, F1-score only 0.8754.
YOLOv3 algorithm because the original NMS method is used in the post-processing process to do IoU calculation, in the dense target scene will produce the detection box overlap problem, if the threshold setting is too large, too many deleted detection boxes will reduce the accuracy. So this chapter uses this as an optimization point to further improve the algorithm.

(a)YOLOv3

(b)SE-YOLOv3

Figure 5. Test set results

5. Summary
This paper is mainly aimed at the problem of too concentrated stacking of prediction boxes in a dense disease scenario where bridge disease data is concentrated. Based on YOLOv3, a new feature layer branch soft-IoU and a new clustering EM-Merger unit are introduced to build a new network model of SE-YOLOv3. Comparative experiments on the self-made bridge disease data set show that the model can not only improve the accuracy of disease in dense scenes, but the detection speed is not significantly improved, and there is still much room for improvement.

Acknowledgments
As the dissertation is about to be completed, I would like to express my most sincere thanks and best wishes to everyone who has helped me.

First of all, I would like to thank my mentor, Associate Professor Zhang Zhi, for the topic selection, data collection, and algorithm experiments. Teacher Zhang has always given me detailed guidance to help me develop ideas. With the teacher’s enthusiasm and continuous encouragement, I was able to Successful completion of graduation thesis.
Secondly, I would like to thank Professor Gu Jinguang of the laboratory. Teacher Gu has given me many opportunities to exercise for large-scale company projects, improved my programming skills and multiple analytical ideas, and also helped me a lot in learning and scientific research.

Next, I want to thank my classmates and roommates in the laboratory for their help and care in study and life, which made my graduate career happy and fulfilling.

Then, I want to thank my family and friends for their continued support and encouragement, and for giving me the courage and strength to move forward.

Finally, I would like to thank all the experts and professors who participated in the thesis review, and thank you for your valuable advice and enthusiastic guidance during your busy schedule.

References
[1] Redmon, Joseph, Farhadi, Ali. YOLOv3: An Incremental Improvement[J].
[2] Goldman E, Herzig R, Eisenschtat A, et al. Precise Detection in Densely Packed Scenes[C]/ 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[3] Ren, Shaoqing, He, Kaiming, Girshick, Ross, 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.
[4] Pascal Pons, Matthieu Latapy. Post-Processing Hierarchical Community Structures: Quality Improvements and Multi-scale View[J]. Theoretical Computer Science, 2011, 412(8-10):892-900.
[5] Bodla N, Singh B, Chellappa R, et al. Soft-NMS -- Improving Object Detection With One Line of Code[J]. 2017.
[6] Jiang B, Luo R, Mao J, et al. Acquisition of Localization Confidence for Accurate Object Detection[J]. 2018.
[7] Yu Aihua. Parameter Estimation of Gaussian Hybrid Model Based on EM Algorithm [J]. Modern Computer (Professional Edition), 2011(17):3-7.
[8] Wang Zhaojun. Necessary condition for convergence of EM algorithm [J]. Journal of Nankai University: Natural Science Edition (Issue 2):85-88.
[9] Zhang Z, Torr P H S. Object Proposal Generation Using Two-Stage Cascade SVMs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1):102-115.
[10] Zhou X, Zhuo J, Krähenbühl P. Bottom-up Object Detection by Grouping Extreme and Center Points. In: arXiv e-prints. 2019. p. arXiv:1901.08043.