DETECTING PROTRUSION LESION IN DIGESTIVE TRACT USING A SINGLE-STAGE DETECTION METHOD

Liansheng Wang[1], Shuxin Wang[1], Shaohui Huang[1,*], Changhua Liu[2,*]

[1]Department of Computer Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China - (lswang, hsh)@xmu.edu.cn

[2]Department of Medical Imaging, The Chenggong Hospital Affiliated to Xiamen University, Xiamen 361005, China - liuxingc@126.com

Commission II, WG II/10

KEY WORDS: Protrusion Lesion, Deep Learning, Single-stage Method, Multi-scale Feature Layers

ABSTRACT:

The classification networks have already existed for a long time and achieve great success. However, in biomedical image processing, classifying normal and abnormal ones only is not enough clinically, the desired output should include localization, i.e., where the lesion is. In this paper, we present a method for detecting protrusion lesion in digestive tract. We use a deep learning-based model to build a computer-aided diagnosis system to help doctors examine the intestinal diseases. Learn from existing detection method, one-stage and two-stage detection algorithm, a new network suitable for protrusion lesion detection is proposed. We inherit the method of anchor generation in SSD, a fast single-stage object detector outperform R-CNN series in terms of speed. Multi-scale feature layers are assigned to generate different sizes of default anchor boxes. Different from the previous work, our method doesnt require additional preprocessing because the network can learn features autonomously. For the 256*256 input, our method achieves 73% AP, perform a novel way to detect protrusion lesions.

1. INTRODUCTION

Protrusion lesion in the digestive tract. Colorectal cancer is the third most common cancer in the world, which developed from untreated protrusion lesion such as polyps. Detecting protrusion lesion in their early stage has become a serious medical issue. Wireless capsule endoscopy (WCE) was designed to examine the intestinal diseases without surgery and can give a direct visualization of intestines, in which the time cost is to analyze lots of images to capture the abnormal parts. The doctors will be wearied with about 55000 images per person produced by WCE. This study aims to build a computer-aided diagnosis system to help doctors analyze and detect lesions in WCE images, mark where the lesions are.

Object detection has obtained pretty good performance in the last few years, measured on the canonical PASCAL VOC datasets or COCO datasets. Current state-of-the-art object detectors are based on a two-stage mechanism, like (Ren et al., 2015), (Girshick et al., 2014), (Girshick, 2015). The first stage generates anchors and the second stage classifies those anchors as background or foreground, regresses the location of the anchors at the same time. Despite the pretty accuracy of the two-stage detector, the speed is slow. Some one-stage mechanisms are proposed aimed to increase speed such as (Lin et al., 2017), (Liu et al., 2016). Recently, object detection has been widely used in face detection and car detection, but to our limited knowledge, those methods are seldom used in medical issue especially in intestinal diseases detection.

Some efforts have been devoted to classifying the specific lesion image in the digestive tract. Yixuan Yuan et al in (Yuan et al., 2017), (Yuan et al., 2015a), (Yuan et al., 2015b), (Yuan and Meng, 2016), (Yuan and Meng, 2015) proposed a WCE abnormal image detection method based on Saliency, the abnormal image include lesion like bleeding, polyps and ulcers. It is a multi-classification task. In (Yuan et al., 2016), (Yuan and Meng, 2014), the improved bag of feature method is used to extract features and the SVM is applied to classify those features to get abnormal images. All those work are to classify normal and abnormal images using manual features which are insufficient to present the lesion. Michael F Byrne et al. developed an artificial intelligence model for real-time assessment of colorectal polyps images in (Byrne et al., 2017). They build a DCNN model based on the inception network architecture to learn features and minimize a frame-level cross-entropy loss function. The methods mentioned above are classification tasks, but more needed in clinical practice is whether the lesion exists and where it is.

Therefore, the mean goal of this paper is to introduce the object detection algorithm into protrusion lesion detection in the digestive tract. Our purpose is to build a deep network to learn features automatically without any extra preprocessing and then detect the location of the protrusion lesion, including polyps, lymphoid hyperplasias, and submucosal eminence.

2. METHOD

We use a simple single neural network to detect protrusion lesion. Like any other single-stage detection models, our network composes a backbone part and two sub-parts. The backbone network computes convolutional feature from the entire image, then the next two sub-networks predict box location and classification respectively. Multi-scale feature maps are used to generate different sizes of anchors for detection, followed by a non-maximum suppression step to remove unnecessary ones.

2.1 Multi-scale feature maps and anchors

We use multi-scale feature maps adapted from the base network, as shown Fig.1. It allows the model to predict and detect at multi-

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W12-231-2019 | © Authors 2019. CC BY 4.0 License.

ple scales. At each feature map, a set of default anchor boxes are generated, and the generation method is the same as (Liu et al., 2016). There are five different levels of feature maps designed in the model, each with different area scales and ratios. In other words, at every feature map cell, anchor boxes will be produced in different shapes and sizes.

During data processing, we compute IoU (Intersection-over-Union, IoU) between every default anchor box and true box that has given in the dataset. Two thresholds, positive IoU threshold and negative IoU threshold, are set for dividing foreground and background. We match default boxes to any true box with the IoU higher than the positive threshold. Otherwise, the boxes will be regarded as background if the IoU value is lower than the negative threshold with all true boxes. If an anchor boxes (g) matches the true box (g), set the offsets (d) as follows:

$$d_i^w = \frac{g_j^w}{e^{\hat{g}_j^w}}; d_i^h = \frac{g_i^h}{e^{\hat{g}_j^h}} \tag{1}$$

$$d_i^{cx} = g_j^{cx} - d_i^w * \hat{g}_j^{cx}; d_i^{cy} = g_j^{cy} - d_i^h * \hat{g}_j^{cy} \tag{2}$$

2.2 Network

The proposed single-stage method for protrusion detection consists of a base network for feature extraction and two sub-networks for classification and location regression. We try VGG-16 network as a base to extract features from an entire image, and the result is barely satisfactory. VGG-16 architecture used for image classification of which the output to an image is a single class label may not learn enough features for detection, so we build a more elegant model. The base model mainly contains 3 × 3 kernels and 5 × 5 kernels to compute convolutional features for classification and location regression. To obtain more representative features, the network structure is not simply a stack of convolutional layers. We draw on the idea of other effective networks like Resnet (He et al., 2015) and (Ronneberger et al., 2015), combining low-level features with high-level features to assemble a more precise output. Similarly, increasing the width of the network, like GooLeNet, will make feature learning more accurate. Analogously, our model uses two different kernels to learn features and combine them together as the input of the next layer. These are indicated in Fig.1.

The base network consists of the repeated application of 3 × 3 convolutions, or the combination of the 3 × 3 convolutions and 5 × 5 kernels, each followed by a rectified linear unit (ReLU) and batch normalization. In total the network has 25 convolutional layers in the base network. Take the output of the base network as the input of classification subnet and location regression subnet, see Fig.2, the model can obtain the confidence and offset for an anchor. For a feature layer of size m × n with p channels, the classification subnet produces c output values which means the confidence of each class at each of the m × n location. And the location regression subnet predicts offset in the form of [x, y, w, h]. Assuming that k boxes are generated at a given location, the network yields (c+4)kmn outputs totally for a feature map.

2.3 Loss

The purpose of the model is to learn the offset values to make anchors match true boxes as good as possible. We use the loss function similar to SSD.

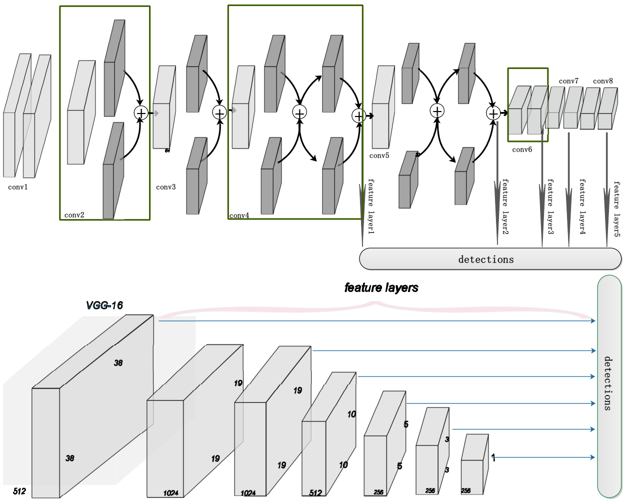


Figure 1. comparison between SSD model and proposed model. Different from SSD, the network structure is not simply a stack of convolutional layers. The idea is by increasing the width of the network in the GooLeNet to make feature learning more accurate. Our model use 3 × 3 kernels and 5 × 5 kernels to learn features and combine them together as the input of the next layer. Both of the two models use several feature layers to get anchors.

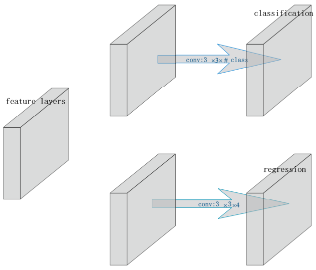


Figure 2. Simple subnet. Take the output of the base network as the input of classification subnet and location regression subnet. After two convolution layers, we get the regression results in the form of [x, y, w, h] and the classification results with confidence.

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \tag{3}$$

Where the L_{conf} means the loss of classification and the L_{loc} means the loss of regression. N is the number of matched default anchor boxes. The confidence loss is the softmax loss and the regression loss is the smooth loss defined in Faster R-CNN [5]. In order to solve the extreme foreground-background class imbalance problem, we calculate the classification loss function of positive and negative samples with a ratio of 1:5, i.e., choosing the negative samples with the greatest loss to achieve the ratio. The regression function only works for anchors with matching true boxes.

$$L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k smooth_{L1}(l_i^m - \hat{g}_j^m) \tag{4}$$

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} log(\hat{c}_i^p) - \sum_{i\in Neg} log(\hat{c}_i^0) \quad (5)$$

3. EXPERIMENTS AND RESULTS

3.1 Dataset and Preprocessing

The dataset used for our model is totally 15000 abnormal images with protrusion lesion from 105 patients, and several authoritative doctors mark the dataset. To train and evaluate the detection model, we divide those data into train, validation, and test set with the percentage of 70%, 10%, 20%. We regard polyps, lymphoid hyperplasias, and submucosal eminence as protrusion lesion and detect the abnormal location in an image. The following images show different numbers and shapes of protrusion lesion in different parts of the digestive tract. The environment of the digestive tract is complicated, so our purpose is to build a stable model to detect lesion as precise as possible. Images are obtained as follows

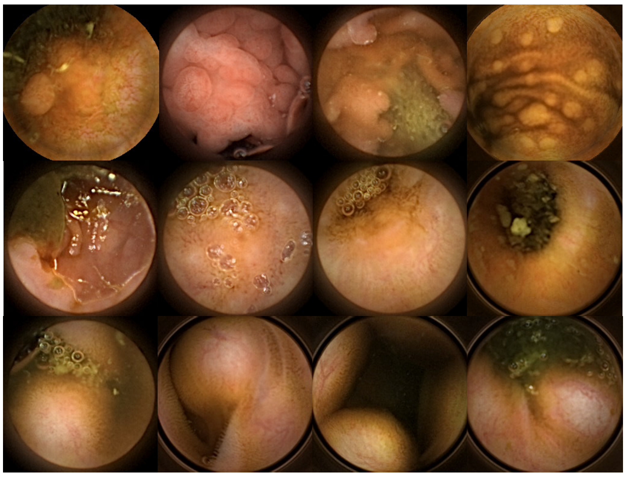


Figure 3. Protrusion lesions of different types in different parts of the digestive tract used for our proposed method. The first row is polyps in different shapes and numbers. The next row is lymphoid hyperplasias and the final is submucosal eminence. All are collectively referred to as protrusion lesions.

3.2 Train and Inference

We use the methods mentioned above to train a model for protrusion lesions detection. The input of the network is 256 × 256. In order to get more dense features, there are five different scales of feature maps which are used for generating anchors. For each feature maps, the base anchor area is 7/256.0, 40/256.0, 80/256.0, 120/256.0, 150/256.0, which adapts to the datasets. Positive IoU threshold of 0.5 and negative IoU threshold of 0.3 are set for dividing foreground anchors and background anchors to train the model. After anchors are produced and features are extracted, the subnet is used to decide whether protrusion lesions exist or not and modify the location of anchors.

During inference time, we obtain the predicted boxes with confidence for a specific class. By using a confidence threshold of 0.6, we can filter out most the less confident. Moreover, to reduce redundancy, we adopt non-maximum suppression (nms) on the boxes to get target results.

3.3 Qualitative Results

In order to train and test the models, approximately 15000 images are prepared. From these, about 11000 images are selected for training, 1500 images for validating, and the rest of them are for testing. We train the model in training images and obtain the result on test images in Fig.4. The blue boxes denote our detection results for protrusion lesions using a threshold of 0.5, and the true boxes are presented in green boxes.

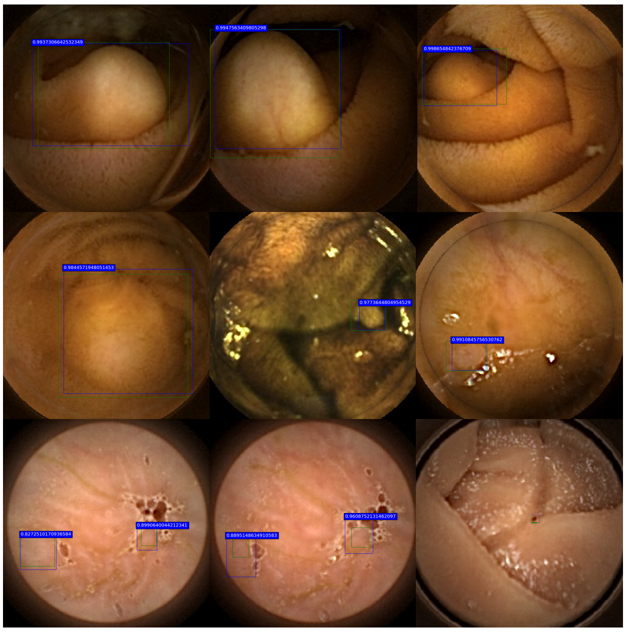


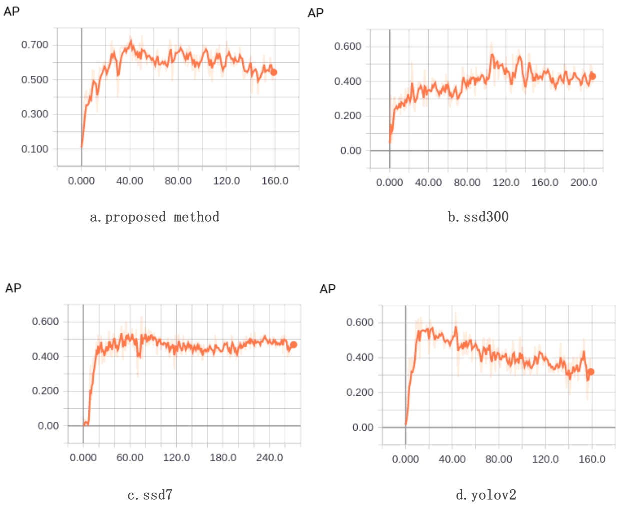
Figure 4. The results of the detection method we proposed. The blue boxes with confidence are our results and the green ones are true boxes.

As shown in Fig.3, there exits different shapes of protrusion lesions. The lesions vary in sizes, ranging from 20× 20 to 150 × 150, and in numbers, too. However, the gastrointestinal environment is complex because of the presence of food debris and villi. A good algorithm needs to overcome all those challenges and detect where the lesions are accurate. Some existing algorithms are not very friendly to small targets. Our method can solve those problems well. As shown in Fig.4, the proposed method can predict protrusion lesions with high confidence and can locate the position accurately. We finally get the average mean of 69.67% at test images, which is a pretty good result in terms of protrusion lesions detection.

3.4 Comparison with Other Detection Methods

We try different networks for our task with the same dataset. We first use SSD model for protrusion detection, but the results are not satisfactory. We guess that it is because of the different sizes of protrusion lesions. In addition, the features extracted by VGG16 in SSD may good for classification tasks, but not good enough for detection. Also, we try Yolov2 with our datasets, the AP curve is computed in validation datasets during training. The quantitative comparison between our proposed method and other approaches is shown in Fig.5.

Fig.5 reports the average mean of different methods on validation dataset during training. It can be observed that the proposed method can achieve better performance than other meth-



a. proposed method

b. ssd300

c. ssd7

d. yolov2

Figure 5. Comparison using our validation dataset for protrusion lesions detection during training: The a,b,c,d curves are for the average mean of our proposed method, SSD300, SSD7 and Yolov2 3/5000 respectively. From the curves, we can see our method performs a novel way to detect protrusion lesions.

ods do, suggesting that our method can generate more representative features to detect protrusion lesions in different shapes and sizes. There are currently two base network architectures of SSD, named SSD300 and SSD7 respectively. The first one, SSD300, is an architecture that is based on a reduced atrous VGG-16, The other, SSD7, is a smaller 7-layer version that can be trained from scratch relatively quickly even on a mid-tier GPU, yet is capable enough to do an OK job on Pascal VOC and a surprisingly good job on datasets with only a few object categories. From the curve we can see, SSD stabilizes at a lower average mean, despite SSD300 or SSD7. Yolov2 get a better performance than SSD at the beginning, but the value of the average mean reduce gradually as the number of training increases. We may have gotten a result of overfitting.

Table 1 shows the average mean results of different methods on the test dataset. Our method achieves the best performance with the average mean. For the protrusion lesions detection, We achieve the AP score of 0.6967 and outperforms Yolov2 which with the best weights saved during training by around 8%, suggesting that the detection method we proposed is a good algorithm for protrusion lesions detection.

Table 1. Comparison with different approaches on our test dateset

Models	feature layers	AP
SSD7	4	0.5687
SSD300	6	0.5764
Yolov2	*	0.6125
Proposed model	5	0.6967

4. CONCLUSION

In this paper, we introduce a detection method, which has been widely used in the computer version, into the medical domain. To the best of our knowledge, few people have applied the detection algorithm to disease detection in the digestive tract so far. However, in clinical practice, physicians need to know the location of the lesion for further observation. Based on this, the model to detect protrusion lesion in the digestive tract is built.

We propose a single-stage object detector reference to SSD. Except for the difference that we modify the network so that it can apply to medical image processing. The model contains a base network and two subnets for classification and box regression. We designed five different scales feature layers from the base network for generating anchors, which have different areas and aspect ratios.

The average precision of protrusion lesions in our method is 73%, which is a pretty good grade comparing with origin SSD and yolov2. We plan to apply this methodology in larger datasets and other medical image analysis problems.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61671399).

REFERENCES

Byrne, M. F., Chapados, N., Soudan, F., Oertel, C., Linares, M. P., Kelly, R., Iqbal, N., Chandelier, F. and Rex, D. K., 2017. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut* pp. gutjnl–2017–314547.

Girshick, R., 2015. Fast r-cnn. *Computer Science*.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Deep residual learning for image recognition. pp. 770–778.

Lin, T. Y., Goyal, P., Girshick, R., He, K. and Dollar, P., 2017. Focal loss for dense object detection. pp. 2999–3007.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. and Berg, A. C., 2016. Ssd: Single shot multibox detector. In: *European Conference on Computer Vision*, pp. 21–37.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In: *International Conference on Neural Information Processing Systems*, pp. 91–99.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241.

Yuan, Y. and Meng, Q. H., 2014. Polyp classification based on bag of features and saliency in wireless capsule endoscopy. In: *IEEE International Conference on Robotics and Automation*, pp. 3930–3935.

Yuan, Y. and Meng, Q. H., 2015. Automatic bleeding frame detection in the wireless capsule endoscopy images. In: *IEEE International Conference on Robotics and Automation*, pp. 1310–1315.

Yuan, Y. and Meng, Q. H., 2016. A novel global and local saliency coding method for polyp recognition in wce videos. In: *Ieee/rsj International Conference on Intelligent Robots and Systems*, pp. 2394–2399.

Yuan, Y., Li, B. and Meng, Q. H., 2015a. Bleeding frame and region detection in the wireless capsule endoscopy video. *IEEE Journal of Biomedical & Health Informatics* 20(2), pp. 624–630.

Yuan, Y., Li, B. and Meng, Q. H., 2016. Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images. *IEEE Transactions on Automation Science & Engineering* 13(2), pp. 529–535.

Yuan, Y., Li, B. and Meng, Q. H., 2017. Wce abnormality detection based on saliency and adaptive locality-constrained linear coding. *IEEE Transactions on Automation Science & Engineering* PP(99), pp. 1–11.

Yuan, Y., Wang, J., Li, B. and Meng, M., 2015b. Saliency based ulcer detection for wireless capsule endoscopy diagnosis. *IEEE Transactions on Medical Imaging* 34(10), pp. 2046.