Improved YOLOv3 Network Combined with LSTM Model and Attention Module for Cerebral Aneurysm Detection

Du Wenjie[1,2] and Wang Yuanjun[2,*]

[1]*Shanghai Zhongqiao Vocational and Technical University, 201514, China*

[2]*University of Shanghai for Science and Technology, 200093, China*

Abstract: Cerebral aneurysm is a kind of cerebrovascular disease, which is mainly diagnosed by reading the MRA slice data to diagnose whether it is suffering from cerebral aneurysm or not, and the medical image detection method based on deep learning can help doctors to improve the detection accuracy and efficiency. Small target detection and the interference of vascular region are the difficulties in cerebral aneurysm detection, which is prone to misdetection or missed detection. Aiming at these problems, we propose an improved method for cerebral aneurysm detection by introducing the LSTM model and the attention module on the basis of the YOLOv3 network, optimizing it in terms of feature processing, time series information construction, weight allocation, etc., using the structure of the YOLOv3 network to achieve effective feature extraction, the regression ability of the LSTM model to construct the time series information among the sliced sequences, and the attention module to assign weights to improve the detection ability of small targets and prevent the interference of blood vessels on the detected targets to improve the detection performance of the network. The experimental results prove the effectiveness of the above improved method which shows significant improvement in the accuracy and anti-interference detection. There is a significant improvement in the detection of cerebral aneurysms, with the precision index reaching 70.8%, an increase of 8.7%, the recall index reaching 76.2%, an increase of 5.0%, and the mAP index reaching 69.9%, an increase of 4.7%, which improves the ability to detect small targets and reduces the interference of blood vessels with the target of detection.

Keywords: YOLOv3, LSTM, Attention mechanisms, Cerebral aneurysm.

1. INTRODUCTION

Cerebrovascular disease is a major fatal and disabling disease, and cerebral aneurysm is a kind of cerebrovascular disease, which is mainly manifested by the protruding wall of intracranial arterial vessels in the human body, and once the aneurysm is ruptured, it is very likely to endanger the safety of life. Cerebral aneurysms occur mostly in middle-aged and old people between 40 and 60 years old, and there are no obvious signs before the onset of the disease. Clinical observation generally shows episodic headache and nerve compression, etc., so diagnosis is more reliable by means of medical images.

Cerebral aneurysm medical imaging diagnostic tools are mainly based on DSA (Digital Subtraction Angiography), CTA (Computed Tomography Angiography) and MRA (Magnetic Resonance Angiography). DSA can be used as the 'gold standard' but has the disadvantages of being invasive, prone to allergy to contrast media, time-consuming and expensive. With the development and improvement of technology, the diagnostic accuracy of MRA and CTA for cerebral aneurysm is close to that of DSA [1]. In general, a case of brain aneurysm patient MRA examination contains 90 to 130 slices of data, experienced doctors to complete a case of patient reading an average of more than fifteen minutes time, if the computer-assisted detection, can improve the accuracy of the data detection and work efficiency, there is an important application of the research value. Deep learning target detection algorithms are mainly divided into two categories: two-stage target detection algorithms and one-stage target detection algorithms, which have their own advantages and disadvantages in detection accuracy and speed. Two-stage target detection algorithms first generate candidate regions and then perform classification and bounding box regression on these regions, such as R-CNN [2], Faster R-CNN [3], etc. One-stage target detection algorithms treat the target detection task as a regression problem and directly perform target detection on the whole image, including YOLO [4] series and SSD [5], etc.

2. CURRENT STATUS OF TARGET DETECTION RESEARCH

2.1. YOLO Algorithms

YOLO algorithm [4] is a neural network model for target detection, proposed by Joseph Redmon in 2016, from the overall point of view, the YOLO algorithm only needs a single CNN network to complete the two tasks of positioning and classification, by obtaining the pixel data of the image to get the coordinates of the target area and classification probability to achieve end-to-end detection and meet the speed requirement of real-time detection. The CNN network in the model splits the input into N×N grids, each of which is responsible for predicting the target whose focal point falls within the grid, the position of the Bounding Box and the confidence score.

*Address correspondence to this author at the University of Shanghai for Science and Technology, 200093, China; E-mail: yjusst@126.com

E-ISSN: 2313-1047/25

YOLOv3 uses the Darknet-53 base network in the feature extraction stage, extensively adopts the residual structure connection in the residual network Res-Net, and introduces the feature pyramid structure to solve the multi-scale problem in detection; it does not use the traditional pooling layer and fully connected layer, and achieves the tensor size transformation by controlling the step size of the convolution kernel, and adopts Leaky Re-LU as the activation function. YOLOv3 inputs a 416×416 image and obtains 13×13, 26×26, and 52×52 feature maps after feature extraction. The bounding boxes in the dataset were clustered and analyzed using the K-means algorithm, and the 9 anchor boxes obtained were divided into 3 groups, which were assigned to the 13×13, 26×26, and 52×52 feature maps, respectively. The feature map was divided into N × N equal-sized grids, each predicting three bounding boxes. The network predicts four coordinate offsets for each cell as: t_x, t_y, t_w, t_h; the coordinates of the upper-left corner of the offset of a particular cell are (c_x, c_y) and the preselected box size of the bounding box is p_w, p_h. Then the predicted coordinates are b_x, b_y, b_w, b_h, and the relational equation that exists between them is (1)-(4):

$$b_x = \sigma\ (t_x)\ + c_x \tag{1}$$

$$b_y = \sigma\ (t_y)\ + c_y \tag{2}$$

$$b_h = p_h e^{t_h} \tag{3}$$

$$b_w = p_w e^{t_w} \tag{4}$$

2.2. LSTM Structure

Recurrent Neural Network (RNN) is a type of recurrent neural network that takes sequence data as input, recurses in the direction of sequence evolution, and all nodes (recurrent units) are connected in a chain fashion [6]. The most important purpose of using RNN is to construct continuity between the data, *i.e*., to allow a contextual environment between the data, where the next moment state depends not only on the inputs, but also on the state of the previous moment, which gives the whole network a certain memory capacity, and therefore RNN is suitable for de-constructing the information between the temporal sequences.

LSTM (long-short term memory), also known as long-short term memory network [7], belongs to a more special RNN structure. In the deep learning training process, LSTM can well avoid the problem of gradient disappearance or gradient explosion [8]. LSTM model is based on the RNN structure introduced memory cell, in each memory cell contains a variety of gating switches: input gate, forget gate, output gate. The gate switches are denoted by i_t, f_t and o_t respectively, which control the transmission of information in the memory cell. The following equations (5)-(10) can be referred to.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{C_t} = \tanh\ (W_c[h_{t-1}, x_t] + b_c) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t} \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = o_t * tanh\ (C_t) \tag{10}$$

Instead of the activation function in the hidden layer of the RNN, state selection in the hidden layer is performed by controlling these gating switches using these gating units. Its various gating switches act as the name suggests, *i.e*., each gating switch selects to input, forget, or output a certain portion of information. The LSTM structure is shown in Figure **1** [9].

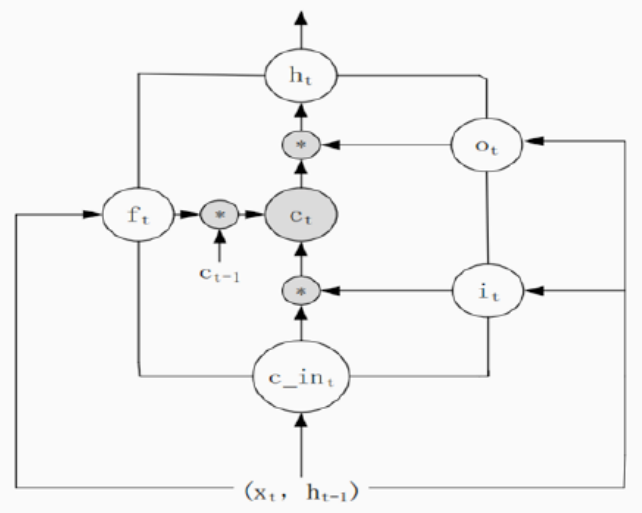

Figure 1: Schematic diagram of LSTM structure.

The first step of the LSTM structure starts with choosing what information can pass through the gating cell, a process controlled by the sigmoid function in the oblivion gate, which generates a value f_t, which is a real number between [0,1], based on the current input x_t and the output h_{t-1} from the previous moment. The f_t value determines which of the learned messages C_{t-1} or C_{t-1} from the previous moment can pass through the gating unit. The second step will generate the updated information, a process where the sigmoid function in the input gate determines what information needs to be updated, generating a new value $\tilde{C_t}$ that relates only to the current layer. The third step is to complete the output of the model, which is controlled by the sigmoid function to control the preliminary output, and then use

the tanh function to transform C_t into a real number between [-1,1], and then multiply the previous two pairs one by one to get the output of the model. LSTM is an excellent model based on the RNN variant with good memory function due to its recursive effect feature. It not only contains the advantages of the RNN structure, but also solves the problem of gradient vanishing or gradient explosion in the backpropagation process. Because LSTM solves the problem with respect to gradients, it is suitable for dealing with long time-dependent sequences of data and data to be processed in an integrated way. These features make LSTM very suitable for sequence related problems.

2.3. Attention Mechanism

Attention mechanisms in the field of deep learning can be interpreted as a means of biasing the allocation of available computational resources towards the most useful parts of a signal [10-12]. At the application level, attention mechanisms are classified into Temporal Attention [13-15] and Spatial Attention [16, 17]; at the level of methods of action, they are classified into Soft Attention [18, 19] and Hard Attention.

The attention function is essentially a mapping relationship that maps a query to a key-value, and the final value vector and its computation is divided into three main steps:

(i) Similarity calculation, calculates the similarity between query and each key to get the weight. Functions used to calculate similarity are dot product, cosine similarity, perceptual machine and so on;

$$X_i = F(q, k_i) \tag{11}$$

(ii) Normalization, the weights are generally normalized using the soft-max function

$$a_i = softmax(s_i) = \frac{exp(s_i)}{\sum_{j=1}^{N} exp(s_j)} \tag{12}$$

(iii) Weighted summation, weighted summation of weights and corresponding key values.

$$Attension\big((Q,K),V\big) = \sum_{i=1}^{N} a_i\, v_i \tag{13}$$

The above processes (i)-(iii) can be described using succinct formulas:

$$Attention\big((Q,K),V\big) = \sum_{i=1}^{N} a_i\, v_i = \sum_{i=1}^{N} \frac{\exp\big(s(k_i,q)\big)}{\sum_j \exp\big(s(k_j,q)\big)} v_i \tag{14}$$

In this paper, the spatial attention mechanism will be chosen to focus on the detection of features of cerebral aneurysms.

3. METHODS

3.1. Difficulties in the Detection

Compared with natural images, generally speaking, the lesion information in medical images is more complex, and most of the brain aneurysms are irregular small targets, relatively speaking, large targets usually contain rich information or obvious features, so it is easier to detect them, but the small targets have the characteristics of low signal-to-noise ratio, small imaging volume, resulting in the inclusion of less useful information, and the lack of detailed features in the target area. In addition, usually in the deep learning model, multiple convolutional layers are used to form a down-sampling layer for image feature extraction, so the learned features will not be very fine, so the precision and recall rate is usually low in small target detection. All of the above reasons lead to the fact that small target brain aneurysm detection in medical images is now one of the research hotspots and difficulties in the field of vision.

Brain aneurysms generally appear in variable locations, shapes, sizes and other information, which makes detection very difficult. Secondly, although some cerebral aneurysms exist independently, they appear on the walls of blood vessels. Since blood vessels and cerebral aneurysms are very similar in terms of brightness, shape, etc., the blood vessels can be very disruptive in the detection process. Therefore, how to distinguish blood vessels from cerebral aneurysms so that the blood vessel portion is not misclassified as a cerebral aneurysm is also the difficulty of this experimental study.

3.2. YOLOv3 Network Combined with LSTM Model and Attention Module

In this paper, we propose an improved YOLOv3 network combined with LSTM model and Attention module, which improves the existing detection method from three different aspects: feature extraction method, construction of temporal information and spatial channel information dependence.

According to the characteristics of the brain aneurysm slice data and the detection task, after introducing the LSTM structure on the basis of the YOLOv3 network, the feature information extracted by YOLOv3, the predicted location information of the target object, and the output information of the previous moment are all inputted into the LSTM model. This structure of LSTM is suitable for the analysis of temporal information, which can extract both the deep feature information of the target object and the temporal information, extending the learning ability of

the network to the spatial and temporal domains. In addition, since the LSTM model has strong regression capability, the detection problem can be transformed into a regression problem, and the output location information is then input into the model, which can play a guiding role for the subsequent detection.

The attention mechanism model is essentially a distribution of weights, with regions that conform to the characteristics of the target object having a greater weight and regions that do not conform having a lesser weight, and the incorporation of spatial attention into the network structure [20] attenuates the interference of blood vessels in the detection of cerebral aneurysms, and this structure of spatial attention allows the network to be more focused on the characteristics of the cerebral aneurysms, increasing the weight of the regions in which they are located.

Based on the characteristics of the brain aneurysm slice data and the detection task, this paper proposes an improved detection method that introduces the LSTM model to extract the visual features and predict the location of the target object with the YOLOv3 network, and then exploits the characteristics of the LSTM model to extend the network's learning and analysis capabilities to both the spatial and temporal domains. The flow of the improved network structure is as follows: firstly, the visual feature information is extracted using the deep CNN structure of YOLOv3, and at the same time, the initial judgement of the target object position is generated, and then the feature information, the position information, and the output information of the previous moment are all fed into the LSTM structure, and finally, the prediction is completed by the network, and the flow is shown in Figure **2**. Adding the LSTM model does not increase the complexity of the network much, and it is trained using the mean square error during the training process, referring to equation (15).

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{b} \| B_{target} - B_{pred} \|_2^2 \quad (15)$$

where n is the number of samples for batch training, B_{target} is the target value of the model, B_{pred} is the predicted value of the model, $\| \cdot \|$ is the Euclidean parameter.

The LSTM model is suitable for temporal processing and also has strong regression capability, which is manifested in two aspects: (i) regression between visual feature information and location information, LSTM as a connecting unit can infer the location information from the visual feature information, which helps to obtain more accurate location information in the following; (ii) cascade linear regression on feature sequence data, the input feature information of current sequence depends on the output prediction information of the previous sequence, which makes the correlation features between consecutive sequences form.

Throughout the fusion model, the input to the LSTM model consists of feature information from the convolutional layer and detection information from the fully connected layer. The formation of the posterior candidate frames is influenced by the regression of the preceding LSTM, which plays a role similar to that of a guide: (i) while the LSTM interprets the input visual features, it also regresses the features to the predicted positions of certain elements; and (ii) the learnt sequence unit of the LSTM restricts the positional predictions to a certain spatial range.

4. EXPERIMENTS

4.1. Data Sources and Pre-Processing

In this experiment, intracranial slice data from magnetic resonance angiography was used, which was first scanned by the MRA machine to form intracranial three-dimensional data in DICOM format, then sampled

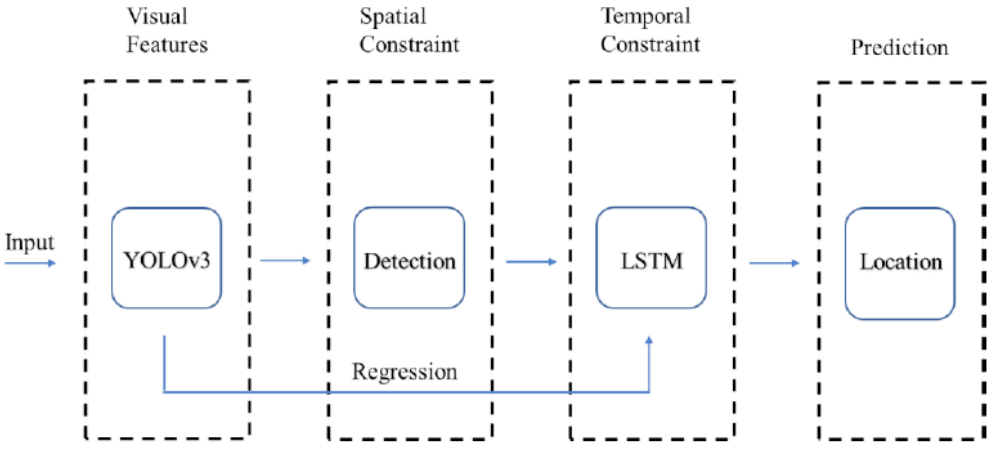

Figure 2: Flowchart incorporating LSTM models.

by the medical software in three directions: transverse, coronal, and sagittal, and finally formed two-dimensional MRA slice data. The number of slices per patient with cerebral aneurysm was not fixed, and most were in the range of 90 to 130. The experimental dataset contained 110 patients (86 males and 24 females, mean age 66 years) from the Concord Hospital with a total of 13840 slices, of which the number of positive samples (*i.e*., slices with cerebral aneurysms in the data) was 1409. The training set and test set were allocated in the ratio of 7:3, the partially sliced data, as shown in Figure **3**.

The process of creating the experimental dataset is as follows: firstly, create and generate the label file in xml format, then convert the xml format file to a txt format file, and then copy the dataset images and files to the corresponding directories, and run to generate the train.txt and test.txt files. The experimental hardware environment is: model i7-7700HQ CPU, model GTX1050Ti 2GB GPU, the software environment is: Window10 operating system and python 3.10 programming language.

4.2. Evaluation Criteria

In order to quantitatively analyze the performance of the improved detection network model in this experiment, the evaluation metrics used in this paper are precision, recall and mean Average Precision (mAP). The precision metric is used to indicate the model's ability to reject non-relevant information and represents the percentage of correct predictions for samples with positive predictions. The recall metric explains the proportion of samples that were actually positive that were correctly judged to be positive. In medical diagnosis, it is necessary to diagnose as many true-positive samples as possible and misclassify as few true-negative samples as possible. The mAP metrics are used to evaluate the classification and localization performance of the model. The formulas for precision, recall, and mAP can be found in (16)-(18), respectively.

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{17}$$

$$mAP = \frac{\sum AveragePrecision}{N(classes)} \tag{18}$$

Where TP (True Positive) means that it is predicted to be a positive sample and the true value is also a positive sample, FN (False Negative) means that it is predicted to be a negative sample but the true value is a positive sample, and FP (False Positive) means that it is predicted to be a positive sample but the true value is a negative sample.

4.3. Experimental Protocols

The batch-size of the training network was set to 16, the weight decay was set to 0.0005, and the momentum was set to 0.9. The learning rate was set at 0.01 for the first 400 rounds of training, 0.001 for the middle 200 rounds, and 0.0001 for the last 100 rounds. The training process is updated using Stochastic Gradient Descent algorithm.

YOLOv3 network will reduce the output feature map to 1/32 of the input, in order to achieve the best

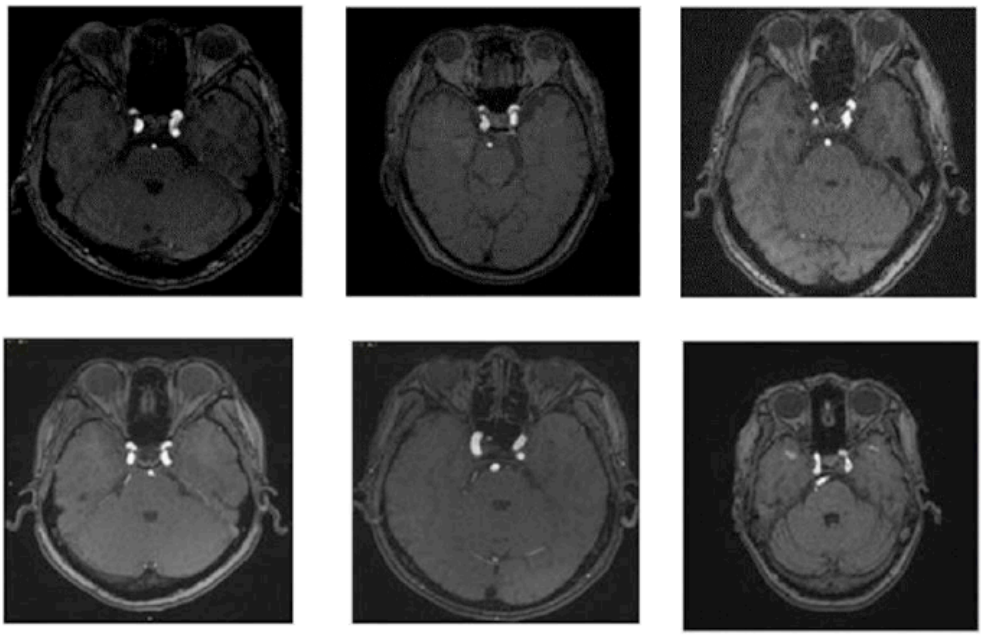

Figure 3: Slicing diagram.

detection effect, the experiment will unify the input image to 416×416 size, and the final size of the output three feature maps are 13×13×31, 26×26×31, 52×52×31. Each grid cell predicts 3 bounding boxes, each bounding box has (x, y, w, h, confidence) five parameters, the experiment only needs to detect the presence or absence of brain aneurysms in one category, the size of the filter is set to 3 × (5 + 1) = 18. Considering the relatively small amount of data in brain aneurysm images, the YOLOv3 network is well suited to be migrated to medical image detection. Instead of training from scratch, the official YOLOv3 pre-trained weights can be used as the starting point for training, and the weights can be fine-tuned before continuing with 700 rounds of training, with all the images in the training set being re-taken as inputs in each round. Such a training strategy will result in better weights for the whole network and its adaptability.

4.4. Experimental Results and Discussion

This paper collates the detection results of the experimental methods on the brain aneurysm slice dataset and performs four comparison experiments, YOLOv3, YOLOv3+LSTM, YOLOv3+Attention, and YOLOv3+LSTM+Attention. The experimental results are shown in Table **1**. As can be seen from the table, the improved method proposed in this paper is indeed able to further improve the detection accuracy based on the YOLOv3 network. The improved network of YOLOv3+LSTM+Attention has a precision metric of 70.8%, a recall metric of 76.2%, and a mAP metric of 69.9%, which is the highest among all the experimental schemes.

Combining the above experimental results, it can be seen from Table **1** that the evaluation indexes are not very high if the YOLOv3 network is used solely for detection. Using the interpretation and regression capabilities of the LSTM model, we were able to extract and construct inter-slice associations to help predict the location of brain aneurysms. The YOLOv3 network combined with the LSTM model resulted in a 5.8% improvement in precision, a 3.4% improvement in recall, and a 3.9% improvement in mAP. Most cerebral aneurysms can be detected accurately for independently existing cerebral aneurysms, but the detection of cerebral aneurysms in the vessel wall is not as good.

When detecting cerebral aneurysms with small targets or cerebral aneurysms attached to the vessel wall, the detection effect needs to continue to be improved, especially for the interference of blood vessels. If the attention module is added, blood vessels and brain aneurysms can be distinguished by increasing the weight of the brain aneurysm portion that is effective for the detection task and decreasing the weight of the blood vessel portion that is ineffective for the detection task.

After combining the LSTM model and Attention module in YOLOv3 network, the network performance is significantly improved and is the most effective in the detection of cerebral aneurysms, with the precision metrics reaching 70.8%, recall metrics reaching 76.2%, and mAP metrics reaching 69.9%, which is 8.7%, 5.0%, and 4.7% higher than that of the original YOLOv3 network. For independently existing cerebral aneurysms, regardless of their size, they can be detected accurately in general. For cerebral aneurysms attached to the vessel wall, there are still few cases of non-detection, but the improvement over the original YOLOv3 network is more obvious.

The combination of YOLOv3 with the LSTM model improves network performance because the LSTM model is able to efficiently regress spatial information from successive different locations of the cerebral arteries. The combination of YOLOv3 with the Attention module improves network performance because the module enhances the acceptance range of the feature map, which leads to more comprehensive information learnt by the network. YOLOv3 combined with Attention module can reduce the interference of blood vessels on the detection because the Attention module can add the weight of cerebral aneurysm and reduce the weight of blood vessels, so as to distinguish between blood vessels and cerebral aneurysms.

5. CONCLUSIONS AND LIMITS

According to the difficulties of cerebral aneurysm detection and the shortcomings of existing methods,

Table 1: Experimental Results of Cerebral Aneurysm Section Data

Method	Precision	Recall	mAP
YOLOv3	62.1%	71.2%	65.2%
YOLOv3+LSTM	67.9%	74.6%	69.1%
YOLOv3+Attention	65.1%	74.1%	66.7%
YOLOv3+LSTM+Attention	70.8%	76.2%	69.9%

this paper proposes an improved target detection method of YOLOv3 network combined with LSTM model and Attention module. Optimization is carried out in terms of feature processing, timing information construction and weight allocation, using the YOLOv3 network structure to achieve effective feature extraction, constructing timing information between slice sequences through the LSTM model, and adopting the Attention Mechanism Module to improve the ability of detecting small targets and preventing blood vessels from interfering with the detection target, to comprehensively improve the accuracy of detection.

The experimental results prove the effectiveness of the improved method proposed in this paper, which shows significant improvement in the accuracy and anti-interference detection for cerebral aneurysm detection. However, for cerebral aneurysms attached to the vessel wall, there is still a problem of missed detection due to the fact that the vessels are extremely similar to cerebral aneurysms in terms of location, brightness and other characteristic information.

The LSTM model and attention module used in this paper is a relatively simple mechanism, in order to carry out more detailed research of target detection in the future, it is recommended to construct a more powerful and stable model, and to make the whole network structure not too complex.

CONFLICTS OF INTEREST

The authors report no conflict of interest, and this research received no specific grant from any funding agency.

The authors wrote the article without using artificial intelligence.

REFERENCE

[1] CHEN Meng, GENG Chen, LI Yu-xin, *et al*. Automatic Detection for Cerebral Aneurysms in TOF-MRA Images Based on Fuzzy Label and Deep Learning [J]. Chinese Journal of Magnetic Resonance. 2022; 39(3): 267-277.

[2] Girshick, Ross, Donahue, Jeff, Darrell, Trevor, *et al*. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition: 2014 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), 23-28 June 2014, Columbus, Ohio.: Institute of Electrical and Electronics Engineers, 2014: 580-587. https://doi.org/10.1109/CVPR.2014.81

[3] Mingjian Zhu. Dynamic Feature Pyramid Networks for Object Detection[C]//Fifteenth International Conference on Signal Processing Systems (ICSPS 2023): 17-19 November 2023. Xi an, China. 2024: 130911N.1-130911N.9.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, *et al*. You Only Look Once: Unified, Real-Time Object Detection[C]//29th IEEE Conference on Computer Vision and Pattern Recognition: 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 26 June – 1 July 2016, Las Vegas, Nevada.: Institute of Electrical and Electronics Engineers, 2016: 779-788. https://doi.org/10.1109/CVPR.2016.91

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, *et al*. SSD: Single Shot Multi-Box Detector[C]//Computer vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, p.I.: Springer, 2016: 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[6] Dan Yang, Lichun Yang, Dabiao Zhou. Stripe removal method for remote sensing images based on multi-scale variation model[C]//2019 IEEE International Conference on Signal Processing, Communications and Computing: ICSPCC 2019, Dalian, China, 20-22 September 2019.: Institute of Electrical and Electronics Engineers, 2019: 61-65. https://doi.org/10.1109/ICSPCC46631.2019.8960737

[7] Mingxue Bi, Bingjie Hu, Handong Yu, *et al*. Long and short-term memory neural network multicomponent gas quantification correction algorithm based on sparrow search algorithm[C]//International Conference on Mechatronic Engineering and Artificial Intelligence (MEAI 2023), Part Two of Two Parts: 15-17 December 2023. Shenyang, China. 2024: 130712W.1-130712W.7.

[8] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[9] Bo Jin. Detection of Cerebral Aneurysms Based on Deep Learning. Huazhong University of Science and Technology, 2020:19.

[10] Hugo Larochelle, Geoffrey Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine[C]//Advances in Neural Information Processing Systems 23. vol. 2.: Neural Information Processing Systems, 2010: 1243-1251.

[11] Artsiom Ablavatski, Shijian Lu, Jianfei Cai. Enriched Deep Recurrent Visual Attention Model for Multiple Object Recognition[C]//2017 IEEE Winter Conference on Applications of Computer Vision: [Volume 2 of 2] Pages 660-1314: IEEE Computer Society, 2017: 971-978. https://doi.org/10.1109/WACV.2017.113

[12] Chunxi Wang, Maoshen Jia, Meiran Li, *et al*. Attention is All You Need for Blind Room Volume Estimation[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), Vol.3: Seoul, South Korea.14-19 April 2024. 2024: 1341-1345. https://doi.org/10.1109/ICASSP48485.2024.10447723

[13] Sheng-Hua Zhong, Yan Liu, Feifei Ren, *et al*. Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling[C]//Proceedings of the twenty-seventh AAAI conference on artificial intelligence and the twenty-fifth innovative applications of artificial intelligence conference: 14-18 July 2013, Bellevue, Washington, USA, v.2.: AAAI Press, 2013: 1063-1069. https://doi.org/10.1609/aaai.v27i1.8642

[14] Tao Shen, Jing Jiang, Tianyi Zhou, *et al*. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence: Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth Symposium on Educational Advances in Artificial Intelligence: New Orleans, Louisiana, USA, 2-7 February 2018, Volume Seven.: AAAI Press, 2018: 5446-5455.

[15] Jingkuan Song, Lianli Gao, Zhao Guo, *et al*. Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning[C]//International Joint Conferences on Artificial Intelligence: IJCAI 2017, Melbourne, Australia, 19-25 August 2017, Volume 3, Part A.: Curran Associates, Inc., 2019: 2737-2743. https://doi.org/10.24963/ijcai.2017/381

[16] Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition: 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 26 June – 1 July 2016, Las Vegas, Nevada.: Institute of Electrical and Electronics Engineers, 2016: 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[17] Buracas GT, Boynton GM. The effect of spatial attention on contrast response functions in human visual cortex [J]. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 2007; 27(1): 93-97. https://doi.org/10.1523/JNEUROSCI.3162-06.2007

[18] Yao L, Torabi A, Cho K, *et al*. Video Description Generation Incorporating Spatio-Temporal Features and a Soft-Attention Mechanism [J]. Eprint Arxiv, 2015; 53: 199-211.

[19] Tao Shen, Tianyi Zhou, Guodong Long, *et al*. Reinforced Self-Attention Network: A Hybrid of Hard and Soft Attention for Sequence Modeling[C]//27th International Joint Conference on Artificial Intelligence and 23rd European Conference on Artificial Intelligence: IJCAI-ECAI 2018, Stockholm, Sweden, 13-19 July 2018, Volume 6 of 8: Curran Associates, Inc., 2018: 4345-4352. https://doi.org/10.24963/ijcai.2018/604

[20] David Dembinsky, Fatemeh Azimi, Federico Raue, *et al*. Sequential Spatial Transformer Networks for Salient Object Classification[C]//12th International Conference on Pattern Recognition Applications and Methods: ICPRAM 2023, Lisbon, Portugal, 22-24 February 2023, Part 1 of 2. 2023: 328-335. https://doi.org/10.5220/0011667100003411

Received on 10-01-2025 Accepted on 12-02-2025 Published on 15-02-2025

https://doi.org/10.12974/2313-1047.2025.12.01