Design and implementation of deep learning algorithm for image discrimination of light field confocal endoscopy for esophageal cancer and gastric cancer

Yanyan Tang[1,*], *Tong* Liu[1], *Shigang* Ding[2], *Ye* Wang[2], *Shuahua* Yue[3], *Xuefang* Yang[3], *Xiaohui* Liu[1], and *Li* Lu[1]

[1]Technology and Culture Department, Beijing Computer Center, 100094, Beijing, China
[2]Gastroenterology Department, Peking University Third Hospital, 100191, Beijing, China
[3]Institute of Medical Photonics, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, 100083, Beijing, China

Abstract. The survival rate of early gastric cancer and esophageal cancer is more than 90%. Confocal endoscopy can detect cell morphology and mucosal glandular structure (depth about 500 μm), presenting a cross-sectional microscopic image unfamiliar to both endoscopists and pathologists. Therefore, using computer-aided diagnosis technology to complete real-time artificial intelligence diagnosis of early esophageal and gastric cancer is of great significance for early detection and early treatment of cancer patients. ResNet convolution neural network based image classification model, the introduction of attention mechanism based on CBAM module to improve the performance of the model, to achieve intelligent diagnosis of confocal endoscopy images.

1 Introduction

Esophageal cancer and gastric cancer are the top five types of gastrointestinal cancer in terms of morbidity and mortality. Early and accurate diagnosis is the most effective way to prevent and cure malignant tumors. Studies have proved that the survival rate of early gastric cancer and esophageal cancer can even reach more than 90%.

3D confocal endoscopy is a kind of confocal endoscopy technology which integrates fast, deep, high resolution and 3D imaging. It can provide an accurate basis for the selection of treatment plan by detecting the depth of tumor invasion.3d confocal endoscopy presents transverse microscopic images that are unfamiliar to both endoscopists and pathologists. Beginners need long-term training and cooperation with pathologists to master the use of confocal endoscopy, resulting in a long learning curve. The depth of three-dimensional confocal endoscopy imaging was 500 μm, which was enough to detect the cell morphology and mucosal glandular structure (about 500 μm), thus providing an important diagnostic basis for histopathological examination and judging the depth of early cancer

[*] Corresponding author:tangyy@bcc.ac.cn

invasion. Horizontal resolution of 2.42 μm, with cell-level resolution, can realize the visualization of abnormal cell structure, so as to achieve accurate early diagnosis.

Therefore, using computer-aided diagnosis technology to complete real-time artificial intelligence diagnosis of early esophageal and gastric cancer is of great significance for doctors to reduce the burden and increase the efficiency and early detection and early treatment of cancer patients.

2 The overall design

The overall design process is shown in Fig.1., including the training stage of the convolutional neural network model and the test application stage of the intelligent auxiliary diagnosis system.

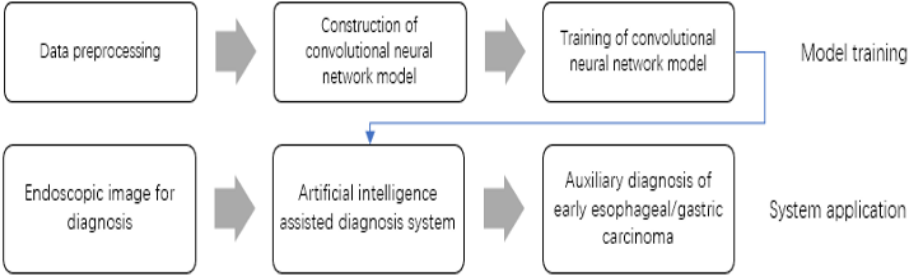

Data preprocessing
Construction of convolutional neural network model
Training of convolutional neural network model
Model training
Endoscopic image for diagnosis
Artificial intelligence assisted diagnosis system
Auxiliary diagnosis of early esophageal/gastric carcinoma
System application

Fig. 1. Design process.

1) In the training stage of the model, the light field confocal endoscopy images provided by The Third Hospital of Peking University and Beihang University should be preprocessed first to realize data normalization and expand the sample size through data enhancement.

Then, an image classification model based on ResNet convolutional neural network is constructed, and the attention mechanism based on CBAM module is introduced to improve the performance of the model.

The training samples and their labels were sent into the classification model, and the network model was trained by setting reasonable loss function and optimization method, so as to obtain the classification model of early esophageal/gastric cancer diagnosis.

2) In the stages of testing and practical application, the light field confocal endoscopy image to be diagnosed only needs to be sent into the trained AI-assisted diagnosis model. Combined with the visual interface of the system, the classification results of the samples to be predicted can be intuitively obtained.

Based on the attention mechanism of CBAM module and combined with ResNet convolutional neural network model, feature extraction of endoscopy image is carried out to make the network focus on more important information and the extracted features cover more critical areas of the image to be identified, effectively improving the recognition accuracy of the model.

3 Detailed design

3.1 Establishment of network model based on ResNet50

The network based on residual blocks is composed of convolutional layer, batch processing normalization layer and modified linear unit Relu.ResNet network effectively alleviates the problems of gradient disappearance, gradient explosion and network degradation caused by network layer deepening by learning residuals and introducing identity mapping

3.2 Introduce spatial attention mechanism

In order to further improve the performance of the model, CBAM attention mechanism is added to the convolutional neural network to enhance its expression ability.

The attention mechanism of the convolution module CBAM is an attention mechanism module combining spatial and channel.

Compared with SENet's attention mechanism that only focuses on channels, it can achieve better results.

1) The Channel Attention Module compresses feature maps in spatial dimensions to obtain a one-dimensional vector before operation.

In the compression of spatial dimension, not only Average Pooling but also Max Pooling are considered.

Average pooling and maximum pooling can be used to aggregate the spatial information of feature maps, send it to a shared network, compress the spatial dimension of the input feature maps, and sum and merge element by element to generate channel attention maps.

On a graph alone, channel attention focuses on what is important on the graph. Mean pooling gives feedback to every pixel on the feature graph, while maximum pooling only gives feedback to the place with the largest response in the feature graph during gradient back propagation calculation, and its mathematical expression is as follows:

$$\begin{aligned}\mathrm{M_c}(\mathbf{F}) &= \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(\mathbf{F})\right) + \mathrm{MLP}\left(\mathrm{MaxPool}(\mathbf{F})\right)\right)\\ &= \sigma\left(\mathrm{W_1}\left(\mathrm{W_0}\left(\mathbf{F}^{\mathrm{c}}_{\mathrm{avg}}\right)\right) + \mathrm{W_1}\left(\mathrm{W_0}\left(\mathbf{F}^{\mathrm{c}}_{\mathrm{max}}\right)\right)\right)\end{aligned} \tag{1}$$

σ is the sigmoID operation, and RELU is required after W_0.

2) The Spatial Attention Module compresses channels and pools the average value and maximum value in channel dimensions respectively. The operation of maximum pooling is to extract the maximum value on the channel. The extraction times are height multiplied by width. The operation of average pooling is to extract the average value on the channel, and the extraction times are also height multiplied by width. Then, the previously extracted feature graph is combined to obtain a feature graph with channel number of 2, and its mathematical expression is as follows:

$$\begin{aligned}\mathrm{M_s}(\mathbf{F}) &= \sigma\left(f^{7\times7}\left(\left[\mathrm{AvgPool}(\mathbf{F}); \mathrm{MaxPool}(\mathbf{F})\right]\right)\right)\\ &= \sigma\left(f^{7\times7}\left(\left[\boldsymbol{F}^{s}_{avg}; \boldsymbol{F}^{s}_{max}\right]\right)\right)\end{aligned} \tag{2}$$

σ represents the sigmoid operation, and 7*7 represents the size of the convolution kernel.

4 The implementation

4.1 Data set preprocessing

The total number of endoscopic images in the data set of this project is 10,000.Then, about 8000 images were randomly selected to form the training set and 2000 images to form the test set in the ratio of 8:2.The data enhancement methods adopted include image random flip left and right, image contrast adjustment, image random illumination and so on.After the data set was expanded, all the data were randomly clipped to 256×256, and common image transformation methods such as random rotation, random horizontal flip and center clipping were used to process the data in order to eliminate the correlation between different features.

Finally, the data is transformed into tensors and normalized. Normalization can make subsequent data processing more convenient and accelerate the convergence speed of model training.

4.2 Intelligent auxiliary model training based on ResNet

4.2.1 Loss function

The Cross Entropy function is used as the loss function of the model.Because cross entropy involves calculating the probability of each category, it appears almost every time with the Sigmoid or Softmax functions.

In the case of dichotomies, there are only two results that the model needs to predict in the end. For each category, the probability obtained by our prediction is P and 1-P, and the loss function is expressed as follows:

$$\mathrm{L} = \frac{1}{N}\sum_{i} -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{3}$$

y_i represents the label value of sample i, $p_i$is the probability that sample i is predicted to be positive.

4.2.2 Optimization method

In this paper, Adam optimization method is used to train the convolutional neural network model. Adam algorithm iteratively updates the weight of the convolutional neural network based on the training data, and designs independent adaptive learning rates for different parameters by calculating the first and second moment estimates of the gradient. The proposer of Adam algorithm described it as the set of advantages of two kinds of stochastic gradient descent extension, namely:

1) Adaptive gradient algorithm (AdaGrad) retains a learning rate for each parameter to improve performance on sparse gradients;

2) Root mean square propagation (RMSProp) adaptively preserves the learning rate for each parameter based on the mean of the nearest magnitude of the weight gradient.

Therefore, Adam algorithm has the advantages of both AdaGrad and RMSProp algorithm. It not only calculates the adaptive parameter learning rate based on the first-order moment mean, but also makes full use of the second-order moment mean of the gradient.

4.3 Model training process

The training of convolutional neural network can be divided into two stages: forward propagation and back propagation. In the forward propagation phase, data is propagated from a low level to a high level (forward). In the backpropagation stage, the error is obtained by comparing the result of forward propagation with the expected target result, and then the error is transmitted from high level to bottom level (reverse). Specific training process is as follows:

1) Initialize network weights;

https://doi.org/10.1051/itmconf/20224501035

2) The input data is propagated forward through the convolution layer, the lower sampling layer and the full connection layer successively to obtain the output value;

3) Calculate the error between the output value of the network and the target value;

4) Errors are transmitted back to the network through the full connection layer, the lower sampling layer and the convolution layer successively. When the error is less than the set threshold, the training is finished.

5) The weight is updated according to the obtained error, and step 2 is entered.

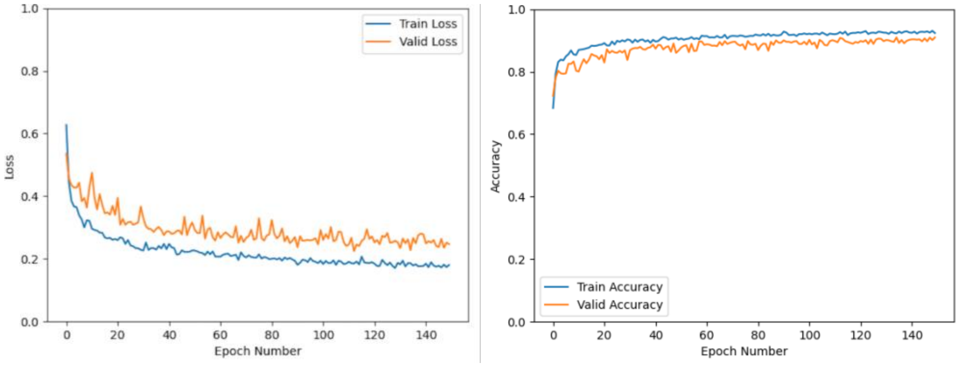


Fig. 2. Loss curve and Acc curve.

Figure 2 shows the change curve of Loss function and prediction accuracy Acc of model prediction results in the training process. Where, learning rate LR =0.003, batch size batCH_size =128, total number of training rounds num_epochs=150.

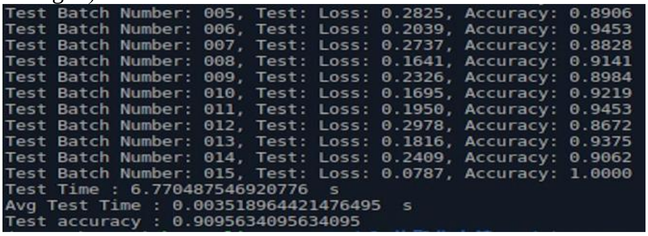
4.4 Test results

Accuracy is used as an evaluation index in this study. Accuracy (ACC) refers to the proportion of all correctly judged results in the total observed values of the classification sample, and its mathematical expression is shown as follows:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where, TP/FP/TN/FN are the four indicators of the confusion matrix: true class TP is the sample that is actually positive class and predicted positive class; False positive FP is the sample that is actually negative and predicted to be positive. The true negative class TN is the sample that is actually negative and predicted to be negative. False negative class FN is the sample that is actually positive and predicted to be negative.

Through the test, the classification accuracy of the model trained in this project reached 90.9563%, and the average prediction time of a single image was 0.0035s (the experimental results are shown in Fig.3.).



```
Test Batch Number: 005, Test: Loss: 0.2825, Accuracy: 0.8906
Test Batch Number: 006, Test: Loss: 0.2039, Accuracy: 0.9453
Test Batch Number: 007, Test: Loss: 0.2737, Accuracy: 0.8828
Test Batch Number: 008, Test: Loss: 0.1641, Accuracy: 0.9141
Test Batch Number: 009, Test: Loss: 0.2326, Accuracy: 0.8984
Test Batch Number: 010, Test: Loss: 0.1695, Accuracy: 0.9219
Test Batch Number: 011, Test: Loss: 0.1950, Accuracy: 0.9453
Test Batch Number: 012, Test: Loss: 0.2978, Accuracy: 0.8672
Test Batch Number: 013, Test: Loss: 0.1816, Accuracy: 0.9375
Test Batch Number: 014, Test: Loss: 0.2409, Accuracy: 0.9062
Test Batch Number: 015, Test: Loss: 0.0787, Accuracy: 1.0000
Test Time : 6.770487546920776  s
Avg Test Time : 0.003518964421476495  s
Test accuracy : 0.9095634095634095
```

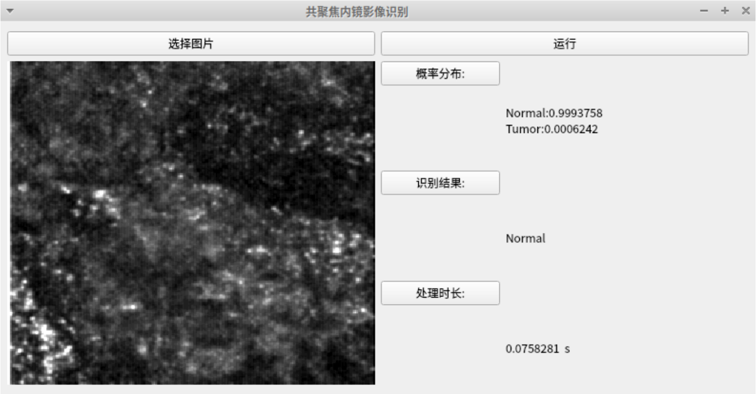
Fig. 3. Model test results.

4.5 Visual interface design.

4.5.1 Initial Status of the home screen

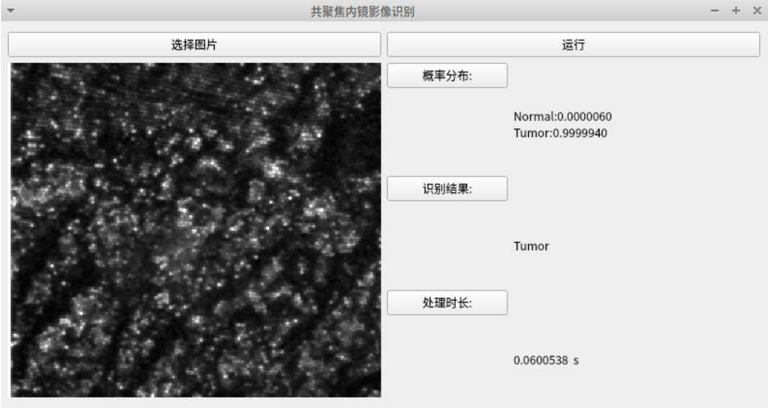
A) "Select Image" button: the image to be predicted under the sample image folder can be selected;

B) "Run" button: the auxiliary diagnostic model can predict the samples, and display the probability distribution of the image (the probability that the sample may be Normal or Tumor), system recognition results, processing time and other information.



```
Normal:0.9993758
Tumor:0.0006242
Normal
0.0758281 s
```

Fig. 4. Display of negative sample test results.



```
Normal:0.0000060
Tumor:0.9999940
Tumor
0.0600538 s
```

Fig. 5. Display of positive sample test results.

5 Conclusion

Multi-mode medical image intelligent auxiliary diagnosis system based on deep learning is realized. The indexes achieved by the system through testing are summarized as follows:

1) The system can classify [normal/cancer] to assist clinical diagnosis;

2) The calculation of a single diagnosis result by the system does not exceed 1 second (average processing of a single image is 0.0035s);

3) The accuracy rate of the system auxiliary diagnosis is no less than 90%;

References

1. Yu He, ZHAO Wenxing. Application of computer aided diagnosis in pathology [J]. Journal of diagnostic pathology,2018,25(03):223-226.
2. Zheng Guangyuan, LIU Xiabi, Han Guanghui. Review of medical image computer-aided detection and diagnosis system [J]. Journal of software,2018,29(05):1471-1514.
3. Shi Jun, Wang Linlin, Wang Shanshan, Chen Yanxia, Wang Gan, Wei Dongming, Liang Shujun, PENG Jialin, Yi Jiajin, Liu Shengfeng, Ni Dong, Wang Mingliang, Zhang Daoqiang, Shen Dinggang. Journal of image and graphics,2020,25(10):1953-1981.]
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.
5. Fei Ling AI, Yuan Ma, Sijia Tian, Xiaonan Wang, Feng Zhang, Xiuhua Guo. Research progress of deep learning in medical image analysis [J]. Beijing biomedical engineering,2018,37(04):433-438.
6. Deng Shasha, Xue Yunjing, Liu Qi, Wang Yousen, Xu Xue, Zhao Xihai, Liu Baiyun. Effects of different CT imaging parameters on the diagnosis of pulmonary nodules with intelligent auxiliary software based on deep learning [J]. Chinese Journal of Medical Imaging, 201,29(10):1003-1006+1011.
7. Zheng Yuanpan, Li Guangyang, Li Ye. Application of deep learning in image recognition [J]. Computer engineering and applications,2019,55(12):20-36.
8. Woo S, Park J , Lee J Y , et al. CBAM: Convolutional Block Attention Module[J]. Springer, Cham, 2018.