Journal of Artificial Intelligence Practice (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2022.050113
ISSN 2371-8412 Vol. 5 Num. 1

Lung Disease Diagnosis based on Transfer Learning

Wanle Chi[1,2*], Yun Huoy Choo[2], Ong Sing Goh[2] and Gong Dafeng[1,2]

[1]*College of artificial intelligence, Wenzhou Polytechnic, Wenzhou, Zhejiang 325035, China*

[2]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Malacca 76100, Malaysia*

chiokchi@163.com

**corresponding author*

Keywords: Lung Disease Diagnosis, LIDC; Lung Image, Transfer Learning, Multi-Tasking Learning, Soft Parameter Sharing;

Abstract: The diagnosis of lung nodules is an important indicator of clinical indications of malignant lung diseases such as lung cancer. Traditionally, doctors read CT lung nodule image to judge the lung disease. The difficulty of doctors' judgement leads to missed diagnosis and misdiagnosis. Through the cooperation of human intelligence and artificial intelligence, computer aided diagnosis can improve the medical imaging clinical diagnosis, improve diagnosis efficiency and accuracy.The objective of paper is to develop a computer-aided diagnostic predictor for lung disease. This paper proposes using cheap and many malignancy labels to transfer learn to expensive and few pathological diagnosis. The paper proposes a partially soft parameter sharing method. In the LIDC datasets, the result of experiment shows that the algorithm of paper is more accurate than other approaches.

1. Introduction

The diagnosis of lung nodules is an important indicator of early clinical indications of malignant lung diseases such as lung cancer[1]. Traditionally, doctors read each CT image of the patient's lung to judge the description and malignancy of lung nodules. The difficulty of doctors' judgement leads to missed diagnosis and misdiagnosis. Through the cooperation of human intelligence and artificial intelligence, computer aided diagnosis can improve the medical imaging clinical diagnosis, improve diagnosis efficiency and accuracy.

Deep learning technology are applied to computer aided diagnosis to realize the automatic diagnosis of disease that achieved good results. Doctors use attributes to describe the shape and appearance of lung nodules. The diagnosis of malignancy has an important reference value for risk assessment of lung disease. The differentiation of benign and malignant lung nodules are the top priority in the auxiliary diagnosis of lung disease.

The accurate diagnosis of lung diseases often requires surgery such as thoracotomy or puncture. Therefore, the diagnosis of lung diseases is costly and harmful to patients. It is easier for doctors to judge the malignancy of lung nodules by reading CT images. And more data can be obtained.

The auto diagnosis of lung images is mainly realized in two approaches. One is an approach

based on the basal learning strategies. It includes feature extraction and kernel methods such as wavelet analysis, non-parametric local texture feature descriptors, decision tree method, and weighted sparse coding method. Han F. proposed Gamb texture, edge features, and non-parametric local texture feature descriptors to train a sparse robust classifier for classifying lung images[2]. The approach cannot automatically extract feature, and difficult to extract subtle features. Another one is the convolutional neural network. The CNN model are outstanding in artificial intelligence such as biological identification[3], human-computer interaction[4,5], and medical image analysis[6]. The approach can process the original images and automatically learn the features to avoid the complexity and limitation of traditional methods.

This paper proposes an improved transfer learning method using the malignancy level of lung nodules images to assist in the diagnosis of lung diseases.

2. Datasets and Preprocession

The datasets are Lung Image Database Consortium-Image Database Resource Initiative (LIDC-IDRI), which consists of medical image files (.dcm) of the chest and corresponding diagnostic lesion markers(.xml). The data were collected at the initiative of the National Cancer Institute (NCI) to study early cancer detection in high-risk populations.

In the datasets, a total of 1012 examples were included. For each example, the images were diagnostically annotated in two stages by four experienced chest doctors. Each doctor independently diagnosed and labeled the patient location, in which three categories were labeled. Then, each doctor independently reviewed the annotations of the other three doctors and gave their final diagnosis. Such two stages annotation can label all images and avoid forced consensus.

The image file is in Dicom format, which is a standard format for medical images. In addition, there are some auxiliary metadata such as image type, image time and other information. A CT image has 512*512 pixels, and each pixel is represented by 2 bytes in the dicom file. For each sample, it can be seen as a three-dimensional matrix D (Z-slicer *X-rows *Y-cols), Z(slicer) represents the number of slices, X(rows) and Y(cols) respectively Indicates the number of rows and columns of the image (default is 512).For example, LIDC-IDRI-0001 is a matrix of 133 * 512 * 512, a total of 133 slices, each size of slice is 512 * 512.

The original datasets are stored in dcm format. But the images as training data input to the network are mostly in normal picture format. It is necessary to convert the dcm format to jpg format. The MicroDicom viewer software was be used to do batch conversions.

A image of lung parenchyma is a low gray connected area, structures are high. The gray values of the segmented area are concentrated and different from unrelated areas. The global threshold method is an iterative method to obtain the final threshold. The algorithm is as follows.

(1) Set the initial value T for the global threshold. The maximum gray value in the image is Tmax and the minimum gray value is Tmin.

$$T=(Tmax+Tmin)/2 \quad (1)$$

(2) Taking T as the threshold, the image is divided into foreground (gray value greater than or equal to T) and background (gray value less than T). And the average gray value TF of foreground and the average gray value TB of background are calculated.

(3) Update threshold T: T = (TF + TB) / 2.

(4) Repeat steps 2 and 3 until T is not changing.

The next task is to remove the bed board and eliminate the holes. Therefore, the circular structural element with radius 5 is used for closed operation, and extracting the maximum connected area can be used for removing the bed board. The connected area, which less than 1000 pixels is gas pipe, should be removed.

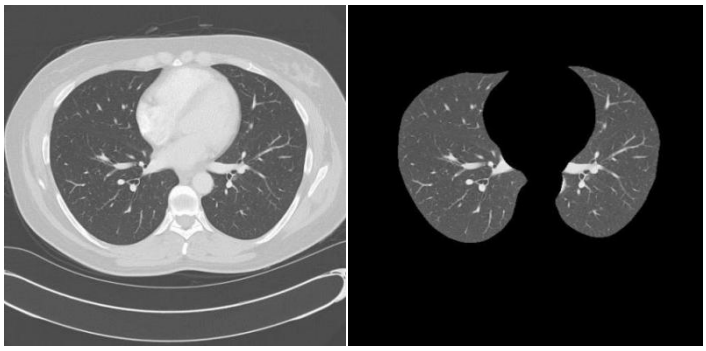

Figure 1: Lung parenchyma segmentation

The label (.xml) file contains information about the doctor's annotation of lung nodules in the CT images, divided into three main categories.

Nodule (3mm-33mm): It contains the characteristics of the nodule and the complete shape of the nodule (roi). The shape information is expressed as the X and Y coordinate value pairs of the nodule polygon in a slice with a certain Z value.

Small-nodule (<3mm): The approximate 3D center of nodule is labeled.

Non-nodule (>3mm): The approximate 3D center is labeled indicating the non-nodule connection area.

The minimum and maximum X and Y of nodules image can be obtained by using the shape information of nodules in the label file, and the rectangular image of nodule is intercepted. Using the OTSU method, an appropriate threshold convert image into a binary image, and areas less than 10 of nodule image were removed. Taking the nodule as the center, black pixels were filled into image to normalize the nodule image into 32 * 32 format. According to the characteristics of 3mm-33mm nodules, the labels show the levels of benign and malignant nodules (level 1-5, 1 is the most benign, 5 is the most malignant). Lung nodule images were extracted and stored with the levels of malignant.

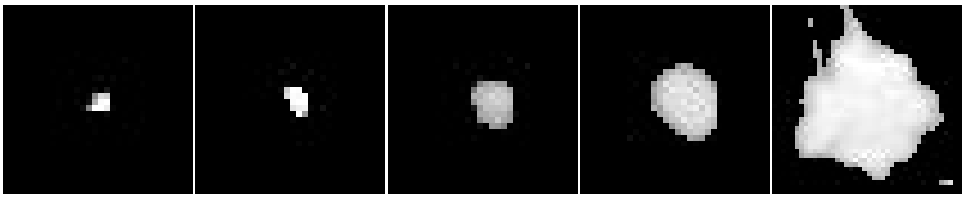

Figure 2: Lung Nodules (From level 1 to level 5)

3. Lung Disease Diagnosis based on Transfer Learning

Pathological diagnosis are difficult to obtain, which requires invasive intervention such as puncture, biopsy, and thoracotomy. Only 300+ samples are marked in LIDC datasets. But pathological diagnosis is accurate for lung diseases diagnosis, and the data are few and costly. The patient needs to endure painful surgical examination.

The diagnosis of nodule malignancy is made by the doctor reading the nodule image, which is cheap and many. LIDC data include1012 examples, more than 249000 images in which 7300 images with obvious nodule and more than 100G data.

The objective of paper is to develop a computer-aided diagnostic predictor for lung disease. Doctors can predict lung disease by reading nodule images. So there are some relationship between pathological diagnosis and malignancy of nodule. This paper proposes using cheap and many

malignancy labels to transfer learn to expensive and few pathological diagnosis.

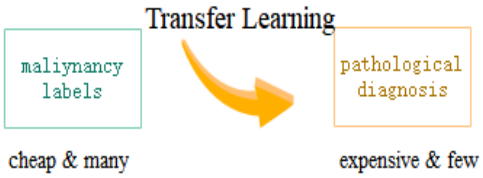

Figure 3: Objective

Multi-tasking learning is a novel method of transfer learning. The multi- tasks learning puts multiple related tasks together to learn at the same time, in which each task has same network framework. Malignancy and pathology learn in same deep learn network, and share parameters at some layers. It can improve the accuracy of malignancy diagnosis.

Multi-task learning method includes hard parameter sharing and soft parameter sharing. In hard parameter sharing method, malignancy classifier over-influences diseases classifier. The over-fitting of disease classifier is reduced, but the accuracy is not high. Soft parameter sharing method reduces the influence and improves the accuracy. This paper propose to reduce sharing more, share only in the pooling layer and the fully connected layer. But the method cannot be degenerated into single task learning. It is a partial soft parameter sharing method.

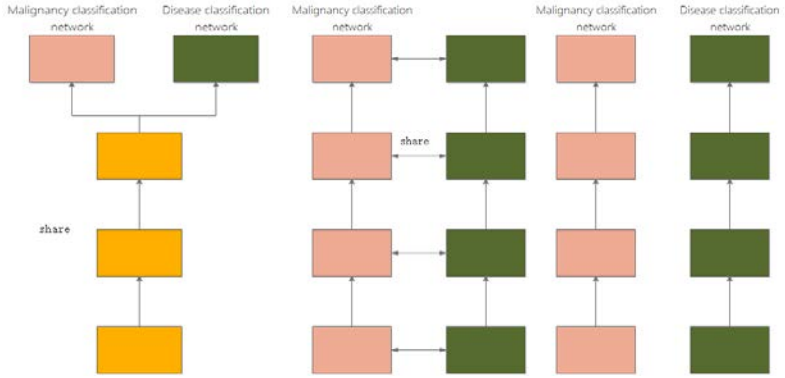

Figure 4: Parameter Sharing, Soft Parameter Sharing and Single Task Learning

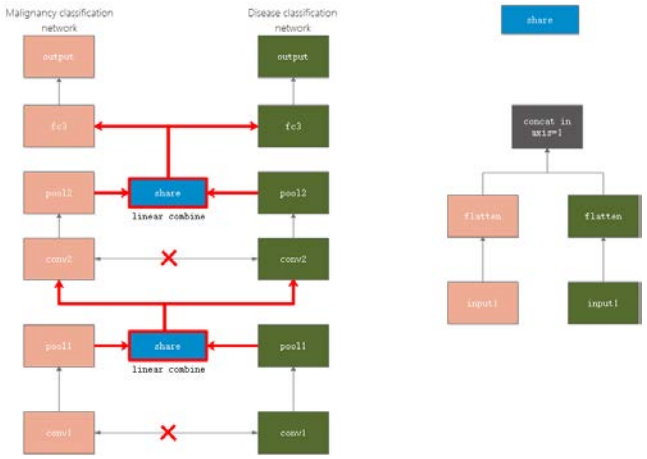

Figure 5: The Structure of partial soft parameter sharing method

The partial soft parameter sharing is used to improve the accuracy of diseases classifier. This structure removed the parameter sharing of conv1 and conv2. The parameters of two tasks are linear combined in pool layers and the full connection layer. The sharing method is to concat and flatten the parameters of two tasks. Use the pathological classifier help to improve the accuracy of diseases malignancy classifier. malignancy data are easier to get and the malignancy classifier is more accurate.

4. Experimental Testing

According to the pathological diagnosis in LIDC datasets, lung diseases were divided into 6 categories, including 0-normal(120 nodules images), 1-normal lung cancer (118 nodules images), 2-non-small cell lung cancer(173 nodules images), 3-pneumonia(138 nodules images), 4-lung lymphoma(126 nodules images) and 5-squamous cell lung cancer(113 nodules images). The experimental environment is shown in the table 1 below.

Table 1: The experiment environment

Name	Value
CPU	1
RAM	16GB
GPU	Tesla v100
VRAM	8GB
Disk	100GB
CUDA	11.0
Python	python3.7
Framework	PaddlePaddle 1.8.0 （TensorFlow）
Calculation package	Paddorch （pytorch）

The experiment divide the nodule datasets into training set and testing set according to 4:1. The experimental result is shown in the table 2, figure 6 and figure 7.

Table 2:The experimental result

global_step_value	epoch	loss_total_value	accuracy_train_value	accuracy_test_value
100	2	3.6882436	0.232	0.2258441
200	4	3.3388233	0.25859375	0.23714285
300	6	3.2287679	0.3164	0.26675325
400	8	3.0206234	0.222	0.293461039
500	10	2.6285636	0.268	0.32097402
600	12	2.3938682	0.27575	0.23493506
700	14	2.2445056	0.3245	0.2762987
800	16	1.763767	0.3328125	0.28116883
900	18	1.7248821	0.35546875	0.29623375
1000	20	1.56265	0.41015625	0.38324675
1100	22	1.2806141	0.43453125	0.3918182
1200	24	1.7929072	0.454703125	0.4283117
1300	26	1.055293	0.56734375	0.45214284
1400	28	0.9944974	0.573203125	0.5614935

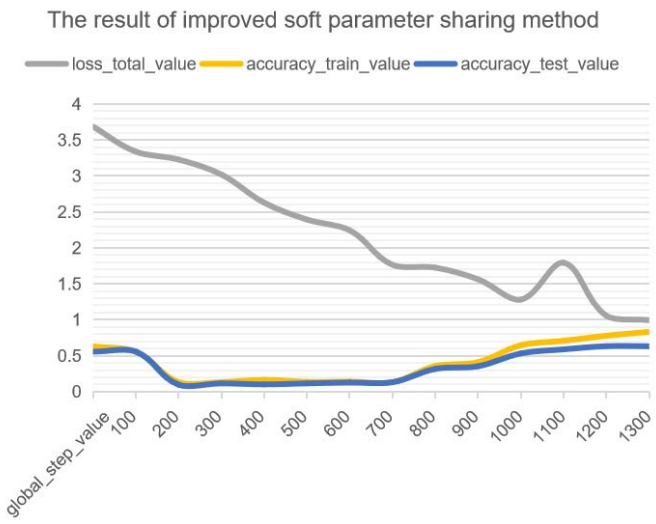

Figure 6: The result of partial soft parameter sharing method

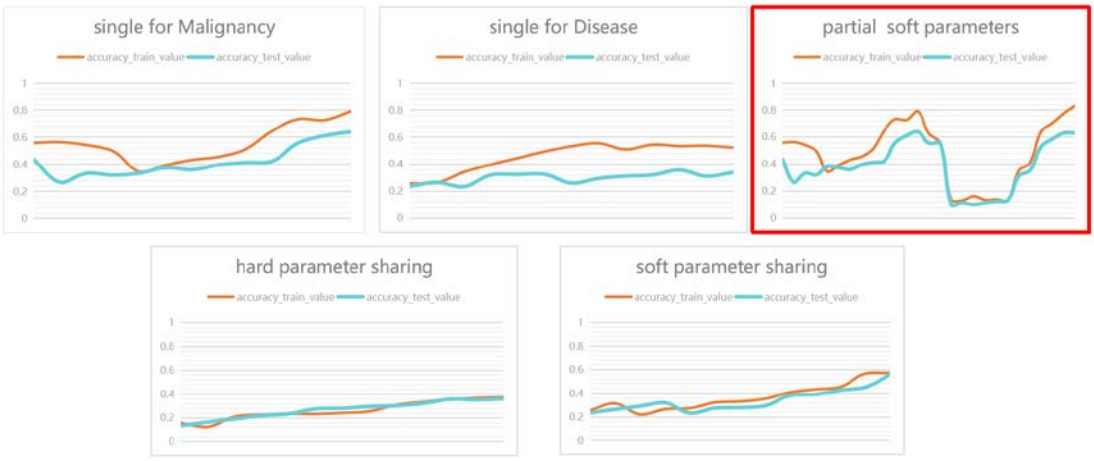

Figure 7: The comparison of algorithm accuracy

The result of experiment shows that the partially soft parameters is more accurate than other methods.

5. Conclusion

This paper proposes using cheap and many malignancy labels to transfer learn to expensive and few pathological diagnosis, to develop a computer-aided lung disease prediction classifier. The paper proposes a partially soft parameter sharing method. In the LIDC datasets, the result of experiment shows that the algorithm of paper is more accurate than other approaches.

Acknowledgements

This work was supported by the Scientific Research Project of Wenzhou Polytechnic, P.R. China (No. XJ2021000101) and the Scientific Research Back Feeding Teaching Project of Wenzhou Polytechnic, P.R. China (No. WZYYFFP2020005).

References

[1] Nie S D,Li - Hong L I,Chen Z X. A CI feature - based lung nodule segmentation using three-domain mean shift clustering. International Conference on Wavelet Analysis and Pattern Recognition. IEEE,2008: 223-227.

[2]Han F, Wang H, Zhang G, et al. Texture feature analysis for computer-aided diagnosis on lung nodules .Journal of Digital Imaging, 2015, 28 (1):99-115.

[3]Zhongqi M, Haosheng Z, Haishi Y , et al. Face expression recognition based on multiple feature fusion dense residual CNN . Computer Application and Software, 2019, 36 (7): 197-201.

[4]Shuo Z, Rong Z. Research on the Handwritten Digital Recognition Algorithm Based on the Convolutional Neural Network Model. Computer Application and Software, 2019, 36 (8): 172-176.

[5]Shun Z, Yihong G, Jinjun W. The development of deep convolutional neural networks and their applications in computer vision. Journal of Computer Science, 2019, 42 (3): 453-482.

[6]Juanxiu T, Guocai L, Shanshan G, et al. Research and Challenges of Deep Learning Methods for Medical Image Analysis. Journal of Automation, 2018, 44 (3): 401-424.