

Title: Intelligent diagnosis on pulmonary nodules

Industrial Applications □Intelligent Manufacturing □Intelligent Driving □Intelligent Life ■Smart Medicine □Smart City

Background description

[Overall background]

Cancer, as a disease that has been hard to overcome by human beings, is a serious threat to human life and health. According to the latest figures published by the National Cancer Center, the incidence of malignant tumors in China in 2013 is 270.59/10 million and the mortality rate is 163.83/10 million, and lung cancer accounts for the first among all malignant tumors and deaths, and about 591 thousand people in China die from lung cancer every year. The survival rate of lung cancer is highly correlated with the stage of disease at the time of first diagnosis. Because early lung cancer has no obvious symptoms, the diagnosis of lung cancer often reaches the middle and late stage. The cost of treatment is high but the effect is not good. Therefore, the early detection and early diagnosis of lung cancer are particularly important. Pulmonary nodules are small, focally, rounded, and denser in shadow. They can be solitary or multiple, without atelectasis, hilar enlargement and pleural effusion. Solitary pulmonary nodules have no typical symptoms. They are usually single, well-defined, denser, less than 3 cm in diameter and surrounded by soft tissue. The popular saying is that there is a white spot in the black lung on CT. The diameter > 3 cm is called lung mass, and the possibility of lung cancer is relatively large. In a recent study of adult chest CT from 2006 to 2012, more than 4 million 800 thousand people had received at least one chest CT examination, found more than 1 million 500 thousand nodules. In 2 years, 63000 patients were diagnosed with new lung cancer. Therefore, although the common small pulmonary nodules are not equal to early lung cancer, the importance of systematic and evidence-based methods to track and monitor these nodules is obvious. At present, the application of artificial intelligence in the medical field has become a trend. Artificial intelligence is applied to the detection of lung nodules and early diagnosis of lung cancer. The lung nodules can be discovered earlier, and then the life of the lung cancer patients can be saved better.

[Business background]

The detection of the larger pulmonary nodules is easy, but the small nodules are relatively difficult to detect and even leak, and the nodules are easily misjudged by the confusion of the images of other tissues such as blood vessels. Using artificial intelligence technology and deep learning algorithm, the diagnosis on pulmonary nodules can be effectively assisted by a doctor with high accuracy, and the detection speed can be increased by dozens of times.

Project description

[Problem description]

According to the chest CT scan images of the patients, the participants need to develop the machine learning algorithm model by themselves, and train the CT scan training data set (MHD format, provided by the organizers) on the labeled and desensitized training data set, so that the trained model can recognize the pulmonary nodules in the CT image intelligently. Hence, it can improve the accuracy of early detection of lung cancer, and reduce the incidence of false-positive misdiagnosis in clinic, so as to achieve early detection, early diagnosis and early treatment.

[User expectations]

The algorithm model designed and developed by the participants can predict the existence of lung nodes in any CT image more accurately and give the prediction probability, and also predict the central position coordinates and radius of the lung nodules to achieve the ability of intelligent identification and intelligent labeling of the pulmonary nodules, and then assist and even replace the doctors to carry out diagnosing on pulmonary nodules.

[Expected economic effect]

It has been set up more than 150 billion Yuan for the whole industry of artificial intelligence, and promoted the development of related industries more than 1 trillion Yuan. In medical health, the data show that the scale of China's medical AI market is also expanding rapidly. It has reached 9 billion 661 million Yuan in 2016, and estimated to exceed 13 billion Yuan in 2017, and is expected to reach 20 billion Yuan in 2018, and the market will continue to expand after that.

Task requirements

[Technical path]

1. Data preprocessing, in view of the medical CT image data provided by the subject, the participants need to deal with them according to the theory or method of image morphology, so that the processed data can better reflect the relevant information needed by the pulmonary nodules prediction and meet the format requirements of the training model for the input data. (Participants can decide for themselves whether to use data enhancement technology).
2. To design the machine learning algorithm model. (it is recommended to use the depth learning algorithm, but the participants can use any type of algorithm) and optimize the model parameters based on the design model, so that the model can predict the size and location information of the nodules in the target CT image as accurately as possible.
- 3, Based on the results of the test set submission model, to submit the CSV file that meets the requirements of the format described below.

[Technical indicators]

The algorithm model of the participants' design training needs to be able to accurately predict whether there is a nodule in the patient's lung area according to the patient's CT image, and to predict the central coordinates of the nodules and the size of the nodule radius: that is to indicate that there are nodules in the sphere represented by the center coordinates and radii of the lung image.

This competition is based on the independent patient CT image test dataset (which is provided by the question party) as the object of evaluation. Its specific evaluation indexes are as follows:

1. The prediction results of all patients' pulmonary nodules provided by the participants, first of all, verify whether each nodule is correct according to the actual results, that is, if the predicted central coordinates of the nodules are within the sphere of the true nodule (radius R), the detection of this nodule is correct.
2. On the basis of the above method, all the nodules predicted by the participants are correct and wrong, and the accuracy of the nodule detection can be obtained by the participants, and a FROC curve is calculated. Finally, the average value of Sensitivity in 1/8, 1/4, 1/2, 1, 2, 4, and 8 is the final evaluation standard value. The higher the value, the higher the score.

[Standard Submission]

In the absence of an external dataset, the contestant completes the nodule prediction test of all patients' CT images on the test set and submits the CSV result file that meets the requirements of the following format. The submission of the CSV file contains 5 columns of data, the first ID value of the patient corresponding to this nodule, second to four for the central coordinates of the nodule with the predictive value of x, y, and z, and fifth as the prediction probability of whether or not the nodule is here. The first line marks the name of each column and marks a detected nodule from each line after the second line. The total number of rows is the total number of nodules of all the patients based on the test data set. The format sample is shown in the "sampleSubmission.csv" file in the dataset.

[Task list]

- (1) CSV result document described above.
- (2) Source code (proposed by Python) is generally included in the following main parts: the image preprocessing module of the original CT image; the algorithm model of the participants and the trained model parameters; the result prediction and output module.
- (3) Contestants' solutions and brief explanations of the reasonableness of the models.

Reference information

[Reference tool]

You can use the SimpleITK package of Python to read MHD format files and use any kind of depth learning framework such as tensorflow, Caffe, keras, CNTK and other deep learning frameworks for model development.

[Reference data]

None

[Data interface]

We used some data in the open data set LIDC/IDRI* as the data of this competition. The use of the dataset complied with the Creative Commons Attribution 3 Unported License protocol.

[*]:<https://wiki.cancerimagingarchive.net/display/Public/LIDCIDRI#2ac66ff98c704d39a7fef84d86578c5a>

training data set download:

<https://pan.baidu.com/s/1jv4id3L2iyZnOOdnSCr0MA>

Password: b5rb,

because of the large data set, is divided into 10 compressed files, corresponding to 10.

Torrent file (you can use Thunderbolt, 115 SkyDrive and other download software to open torrent File, and the annotation file annotations.csv, and submit the example file.

SampleSubmission.csv

Training data set can be downloaded:

<https://pan.baidu.com/s/1jv4id3L2iyZnOOdnSCr0MA>

Password: b5rb

Because the data set is large, it is divided into 10 compressed files, corresponding to 10 torrent

files (which can be used for thunder and 115 nets).

Disk download software to open the torrent file), as well as the tagging file annotations.csv, and submit sample files.

SampleSubmission.csv