Lung Nodule Classification Project

Background introduction:

Lung cancer is the leading cause of cancer-related death worldwide. Lung cancer screening programs using low-dose CT are being implemented in the United States and other countries. Computer-aided detection (CAD) of pulmonary nodules could play an important role when screening is implemented on a large scale.

The Computer Aided Detection System (CAD) is designed to automatically detect pulmonary nodules by computer system based on lung CT scans. Similar to the process of machine learning areas such as object detection, the detection system first extracts some or even a large number of candidates or proposasl, and uses these candidate nodules for the next classification or other tasks .

The goal of the Lung Nodule Classification Project is to classify candidates or proposals for a given nodule to determine whether they are nodules.

Terminology:

|  |  |
| --- | --- |
| **Term** | **Translation** |
| **CT** | CT's full name is computed tomography (computed tomography), the imaging principle is the use of radiation exposure to a certain part of the human body has a certain thickness of the plane, and with the detector to receive the rays. The density of the different parts will affect the amount of radiation penetration. The physical signals received by the detector are reconstructed by the image processing algorithm, and the computer images of each layer are obtained. Appropriate understanding of some knowledge of CT will help you to study the problem, you can refer to Wikipedia and XRayPhysics. |
| **Slice** | The original CT data is reconstructed to form a CT image (2D) of each layer, with each image representing a slice of the lung, called slice. The thickness of the CT images produced by different devices (the actual thickness of each section) is different. For the specific meaning of the thickness, refer to the relevant links given in section CT. However, the relevant parameters of the CT information, including the layer thickness, do not affect your solution to this problem, you can choose to ignore, and we use the selected and unified data. |
| **Scan** | Scan refers to a set of CT images of a subject. For example, a subject's lung CT scan may have more than 300 images, the set of these images is a CT scan. Change  In other words, slice is a two-dimensional image, and a number of slices in order to form a three-dimensional image, that scan. |
| **HU** | HU value is called Hounsfield scale, specifically can see Wikipedia. Briefly, the HU value is a measure of the degree of attenuation of the various parts of the CT. The data we provide are in HU units. However, you can think of a CT image as a single channel or grayscale image, but its range is not [0,255]. |
| **Candidate** | In the field of image processing, for various reasons, the algorithm can only be applied to a local part of the image. Therefore, the algorithm designer will first select some candidate regions on the image to train the classifier or in the algorithm as  Candidates are positive samples, these candidates called candidate, your task is to achieve the classification of these candidates. 🡪 does this mean we first have to train the algorithm on certain regions were the candidates are? |
| Nodule | nodules are solid, elevated areas of tissue or fluid inside or under the skin with a diameter greater than 0.5 centimeters |

🡪 Multiple slices form a scan.

About the data#

Data set

The training set we provided contains 723 scans. For all scans (training and test sets), we have interpolated from the information provided by the data, so that all the images are isotropic, that is, each pixel represents a 1 cubic millimeter of the cube. If you need information about the interpolation or even the original data that has not yet been interpolated, please contact us. We interpolate the CT image, based on the extraction of candidates (the average difference between the positive and negative samples of each scan several times), your goal is to classify these candidates.

Note: This project does not disclose the final test data.

Data mode

We provide the training data to scan as a unit, each scan corresponds to a separate folder, which contains only a file caididates.mat,

Caididates.mat (matlab file) - An array of structures (row vectors) containing all the candidate information for this scan. The fields for each structure (the rows) are:

* CANDIDATE\_ID: the global ID of each candidate; 🡪 so we have one candidate ID per scan? Most likely wrong!
* NODULE\_ID: ID of the current scan in the nodule. Since there may be more than one positive sample corresponding to a true nodule, you can use this ID to select different true nodules. Negative samples have a value of zero. 🡪 one nodule might appear in multiple candidates since the border of the candidates are not exactly at the borders of the nodules, but just somewhere 🡪 so in case I just want distinguished nodules in the training set, I have to distinguish them in the code
* VOL: candidate The HU value of the corresponding 40x40x40 region in the CT image (interpolated to ensure that each voxel represents a cubic millimeter of cube, and the HU value is guaranteed to be non-negative through linear operation). Stored as a single type.
* LABEL: 1 for positive samples and 0 for negative samples.

In addition, we provide a global train\_ground\_truth.mat that you use to calculate your own classifier AUC (code already provided).

Positive and negative sample description

The center of the positive sample is within the true nodule or within 2.5 mm from the true nodule center. Negative samples do not satisfy the above properties, but may contain part of the true nodule.

Evaluation method

Test data format and training data exactly the same, but we will not publish the test data.

The last thing you need to commit is the program code, the algorithm report, the final version of the classifier, the script file that will be able to run the classifier simply by changing the path, and the environment configuration you need to run your classifier (without limitation Framework, but recommended the use of C, C + +, Matlab, Python, Lua and other machine learning framework commonly used language)

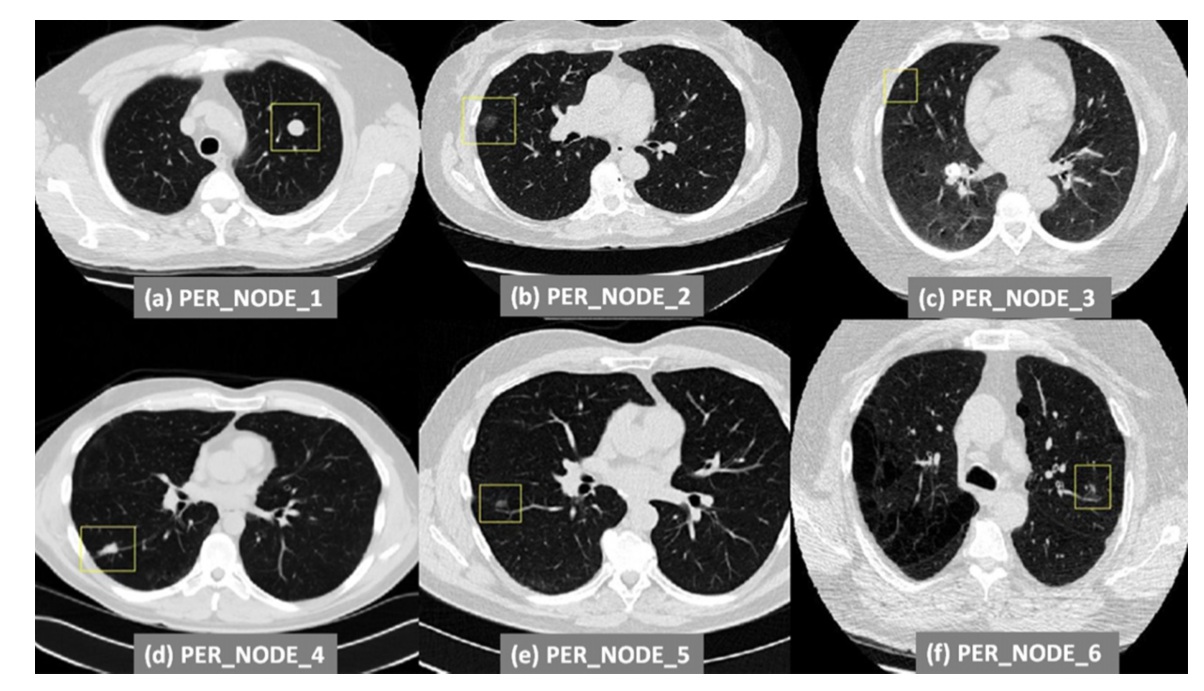
We currently support machine learning frameworks: Caffe, Tensorflow, Keras, Torch, Theano. I suggest you use these frameworks, if you are using a framework other than these Please record your installation process framework, we may need to help you install the framework you use.

You need to run the script is to change the data directory and the output directory of your results can run (you train time train\_data folder, give us the test only need to change the folder path to test\_data). The classifier you submit should output whether or not each candidate is a nodule's confidence. The value is [0,1], and the storage format is mat file. The mat file needs to contain an array of N \* 2, N represents the number of test data, the first column records CANDIDATE\_ID (ascending order), the second column records the confidence, the specific format can refer to our given AUC calculation program. The program evaluates the global AUC for all nodules based on the confidence level.

File List

|  |  |
| --- | --- |
| Data file | Explanation |
| Train\_data | Training data directory, test data format is completely consistent |
| Raw\_data | 10 complete CT images and the corresponding matlab tool |
| AUC.m | Calculate the AUC matlab script |
| Readme.pdf | Documentation |

Nodule knowledge supplement



The picture shows six different types of nodules

A. Isolated large nodules, was isolated spherical, most of the nodules for a similar ball but the size can be changed.

B. Attached to the lung wall of the ground glass nodules, low brightness

C: Pulmonary wall of small nodules, nodules smaller diameter and lower brightness.

D: pulmonary nodules, nodules grow in the lungs (the lung below the nodule that line), the shape is not nearly spherical (cubical)

E: Vessel at the ground-glass nodules, nodules grow in the vicinity of blood vessels (but may have blood vessels through the nodules).

F: vascular nodules.

We provide the raw\_data folder which contains 10 scan full three-dimensional CT image rawVol and pixel-by-pixel marker nodule location rawLabel and the corresponding look at the program. If you feel that it is necessary, you can follow the procedures and procedures to see the input requirements and methods of operation in the scan scan nodules.

Contact persons:

steve\_gao@pku.edu.cn 高俊

gujiayuan@pku.edu.cn 顾家远

1400012710@pku.edu.cn 张梦晓