# Detecting and Locating Lung Tumor Nodules Using Recurrent Attention Model

**Harish Anand**
Senior under-graduate student in Computer Science and Engineering
Model Engineering College, Kerala, India
harishanand.mec@gmail.com
https://github.com/harishanand95/

*Classification of chest Computerized Tomography (CT) scans into malignant, benign and non-nodules cases have always been an interest to medical research community. An expert radiologist detects radiological abnormalities by looking at the "right" locations and making quick comparisons with scans/images from healthy individuals. In this proposal we would like to explore various techniques to assist radiologists in the process of detecting medical conditions from X-rays and CT scans. To that end we propose an implementation and an extension of the paper "Learning what to look in chest X-rays with a recurrent visual attention model" [1]. We aim to apply the idea of the recurrent visual attention model to classify the nodules in chest CT images (or chest X-ray images) using TensorFlow. In particular, we will try to predict the location of nodules and compare it to the nodule location specified in the dataset.*

## 1 Introduction

Lung cancer is one of the most common type of cancer which affects people all over the world. It accounts for 13 % of all new cancer cases and 19 % of cancer related deaths worldwide. About 1.8 million new lung cancer cases were estimated to have occurred in 2012. [2]. Early detection of lung cancer vastly improves the survival chances of the patient and hence, it is of great interest to the medical research community.

A CT scan of a patients chest is usually the first test a radiologist will conduct to check if the patient has a lung cancer or not. Typically tumors more than 3 cm are very likely to be cancer. However, tumors in it early stage can be small and we would like to detect them. Smaller tumors consists of nodules which are growths of size less than 3cm. Larger growths are called masses and are presumed to be malignant. A lung nodule is a small, round growth of tissue within the chest cavity. Nodules can be benign and malignant depending upon size. Large size nodules are considered malignant tumors. We aim to detect the nodules and their location. [3].

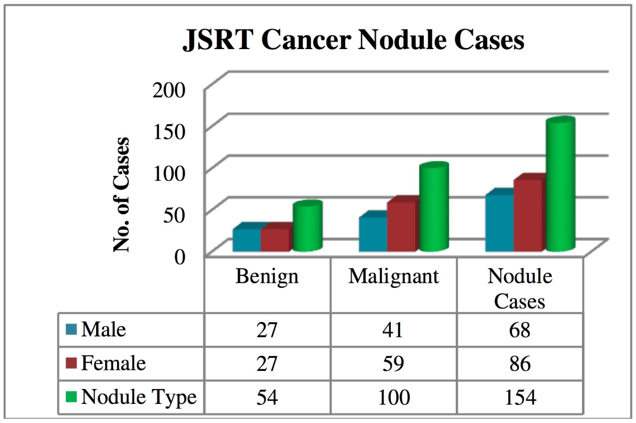
## 2 Related Work

In [1], the authors use a Recurrent Attention Model (RAM) to classify chest X-ray images into normal (i.e., those with no reported abnormalities), an enlarged heart (i.e., cardiomegaly) and a medical device (e.g., a pacemaker). Our proposal is mainly inspired from this paper and we aim to use a similar approach to classify and locate malignant and benign nodules, and non-nodules in lung CT scans. The work [1] uses an attention model which learns to focus and process only in a certain region of an image that is relevant to the classification task. Lung nodule classification is a similar problem and we shall embrace this approach. The JSRT dataset [4] provides the exact location of lung nodules presence and we aim to use it to verify whether the RAM training has reached the desired accuracy in terms of locating the nodule.

The project also takes inspiration from the previous works of converting images to text using convolutional neural network (CNN) and recurrent neural network (RNN), see for e.g., [5, 6] The CNN-RNN model presented in [6] was used for automated image annotation in chest X-rays. The RNNs were trained to describe the context of a detected disease, based on the deep CNN features captured. In [6], the authors were able to generate the disease context clearly, but they had difficulty in trying to locate the disease. For example, cases of "calcified granuloma in *right* upper lobe" and "small calcified granuloma in *left* lung base" were categorized into a general category of "calcified granuloma" or "cardiomegaly". But the doctors/radiologists diagnosis are usually similar to "calcified granuloma in *right* upper lobe" or "small calcified granuloma in *left* lung base" with a clear specification of the size and the location. Our proposed project aims to overcome this drawback by using the attention model to predict the location and possibly the size (large or small) of the nodule.

## 3 Dataset

The standard digital image database of chest lung nodules (JSRT database) was created by the Japanese Society



| | Benign | Malignant | Nodule Cases |
|---|---|---|---|
| 🟦 Male | 27 | 41 | 68 |
| 🟥 Female | 27 | 59 | 86 |
| 🟩 Nodule Type | 54 | 100 | 154 |

Fig. 1. Gender wise distribution of JSRT cancer cases. [7]

of Radiological Technology (JSRT) in cooperation with the Japanese Radiological Society (JRS) in 1998. Database consists of CT images of malignant, benign and non-nodule cases. It has 154 nodule and 93 non-nodule images. Out of the 154 nodule images, 54 are benign and 100 are malignant cases. The images are of high resolution (2048 x 2048 size). The annotations for the images include patient age, gender, diagnosis (malignant or benign), x and y coordinates of nodule, simple description of nodule location, and the degree of subtlety in visual detection of the nodules [4]. The dataset describes nodules in the CT scan by the x and y coordinate of the nodule as well as in text format like *r.upper lobe* or *r.lower lobe*. The data obtained from JSRT database is summarized in Fig. 1.

The dataset in total has only 247 images and is relatively small. There is a shortage of images available for the network to train upon. Inspired by the work [6], we propose two approaches to increase the dataset size.

1. **Scaling and random cropping:** To balance the number of samples for each category, we can augment the training set of the cases with less number of images by randomly cropping 224 x 224 size images from a scaled 256 x 256 size image of the original (2048 x 2048) image. We will be reducing the size of the original image from 2048 x 2048 to 224 x 224 size for inputs. A scaled (to 256 x 256 ) and then cropping to 224 x 224 method makes sure that we lose only the data at the edges.
2. **Adjusting brightness and using horizontal flip:** Changing the brightness of the CT images can be used to get more images as it doesn't affect any spatial representation in the image. Another method is to do a horizontal flip of the image and we must make sure that the nodule location in the description is also changed to the new location in the flipped image description.

## 4 Methods

The RAM model mentioned here is same as the one originally proposed in the paper [1]. Mimicking the human visual attention mechanism, this model learns to focus and process only a certain region of an image that is relevant to the classification task. In [1], this method was used to classify images into categories like enlarged heart, medical devices, and no abnormalities. A similar approach will be used here to detect nodule's locations and classify the images. The model described in the paper [1] is as follows:

1. **Glimpse Layer:** At each time $t$, the model does not have full access to the input image but instead receives a partial observation, or "glimpse", denoted by $x_t$. The glimpse consists of two image patches of different size centered at the same location $l_t$, each one capturing a different context around $l_t$. Both patches are converted to match in size and passed as input to an encoder, as illustrated in Fig. 2.
2. **Encoder:** The encoder implemented here differs from the one used in [8]. The goal of the encoder is to compress the information of the glimpse by extracting a robust representation. To achieve this, each image of the glimpse is passed through a stack of two convolutional autoencoders with max-pooling. Each convolutional autoencoder in the stack is pre-trained separately from the RAM model. During training, at each time t the glimpse representation is concatenated with the location representation and passed as input to a fully connected (FC) layer. The output of the FC layer is denoted as $b_t$ and is passed as input to the core RAM model, as seen in Fig. 2.
3. **Core RAM**: In each time step $t$, the output vector $b_t$ and the previous hidden representation $h_{t-1}$ are passed as input to the LSTM layer. The locator receives the hidden representation $h_t$ from the LSTM unit and passes on to a FC layer, resulting in a vector $o_t$. The locator then decides the position of the next glimpse by sampling $l_{t+1}$ $\sim$N($o_t$, $\Sigma$), i.e. from a normal distribution with mean $o_t$ and diagonal covariance matrix $\Sigma$. The location $l_{t+1}$ represents the x, y coordinates of the glimpse at time step $t+1$. At the very first step, we initiate the algorithm at the center of the image, and always use a fixed variance $\sigma^2$.
   The classification task is done at the last softmax layer on $h_n$ in the nth iteration. Fig. 3 shows the movement of glimpse layer towards the pacemaker in the base paper. [1] The circle and triangle points (in red) indicate the coordinates of the first and last glimpse in the learnt policy, respectively. Fig. 3 (A) shows the locations mostly attended by the RAM model when looking for medical devices. From this figure it is obvious that the learnt policy explores only the relevant areas where these devices can generally be found. Two examples of paths followed by the algorithm after learning the policy are illustrated in Fig. 3 (B), (C). In these examples, starting from the center of the image, the glimpse layer moves closer to a region that is likely to contain a pacemaker, which is then correctly identified. This idea can be adapted to find the location in lung nodules in our problem.
   In our case, an additional locator is needed in the last layer to get the position of the nodules. We can use its

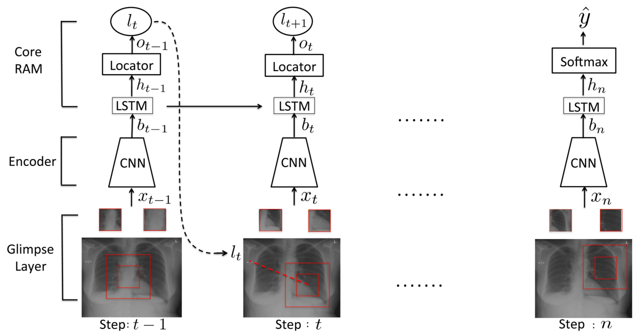


Fig. 2. RAM. At each time step t the Core RAM samples a location $l_{t+1}$ of where to attend next. The location $l_{t+1}$ is used to extract the glimpse (red frames of different size). The image patches are downscaled and passed through the encoder. The representation of the encoder and the previous hidden state $h_{t-1}$ of the Core RAM are passed as inputs to the LSTM of the step t. The locator receives as input the hidden state $h_t$ of the current LSTM and then it samples the location coordinates for the glimpse in the next step t + 1. This process continuous recursively until step n where the output of the LSTM $h_n$ is used to classify the input image

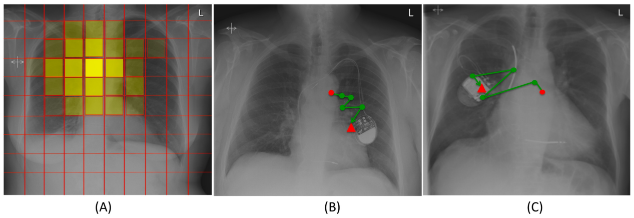


Fig. 3. (A) Image locations attended by the RAM model for the detection of medical devices in the base paper. (B) and (C) are two different samples of the learnt policy on test images in the base paper.

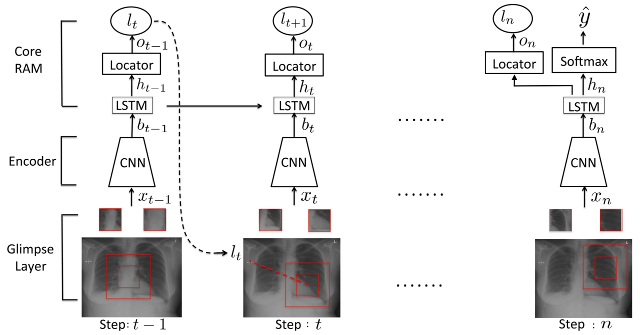


Fig. 4. The slightly modified approach having a locator in the nth step is shown.

output $l_n$ to verify with the nodule location specified in the dataset. The modified approach is shown here in Fig. 4.

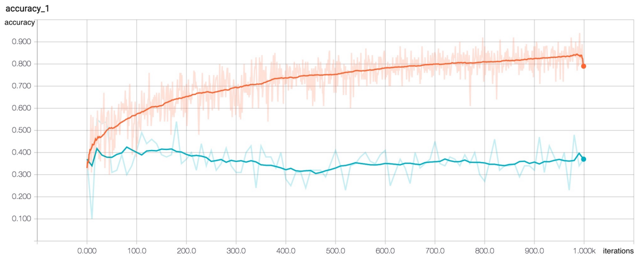


Fig. 5. Retraining inception's final layer for lung nodules.

## 5 Preliminary work

Classification of lung nodules based on an approach of retraining the last layer of the inception was done as a preliminary work [9]. Retraining inception's final layer for new categories is a method mentioned in Tensorflow tutorials. [link]. The work involved only the creation of image categories for training and run the inception network as per the tutorial. The orange line denotes the accuracy for training image dataset and the blue line indicates the validation accuracy. The validation dataset accuracy is poor compared to the training dataset. The training results shows how the network hasn't learned the essential features to look for like the nodules, after 1000 iterations. We aim to overcome these challenges using the RAM model.

As a preliminary study we have tried the classification of China Set-The Shenzhen set - Chest X-ray database images into tuberculosis and non-tuberculosis categories. A 4-layer convolutional network was made in tflearn and later in TensorFlow to classify chest x-ray images and also a data parser was made to fetch and generate random test/train images. [link]. The accuracy obtained for the classification of tuberculosis was 62% for the 4 layer network.

## References

[1] P P Ypsilantis, G. M. Learning what to look in chest X-rays with a recurrent visual attention model. See also URL `https://arxiv.org/abs/1701.06452`, 23 January 2017.

[2] Ferlay J, Soerjomataram I, E. M. D. R. E. S. M. C. e. a. L., 2013. International agency for research on cancer. Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11.

[3] David M. Hansell, Alexander A. Bankier, H. M. T. C. M. N. L. M., and Remy, J., 2008. *Fleischner Society: Glossary of Terms for Thoracic Imaging*. RSNA.

[4] J. Shiraishi, S. Katsuragawa, J. I. T. M. T. K. K.-i. K. M. M. H. F. Y. K., and Doi, K., eds., 2000. *Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists detection of pulmonary nodules*, Vol. **174** of *American Journal of Roentgenology*. pp. 71–74. See also URL `http://www.jsrt.or.jp/jsrt-db/eng.php`.

[5] A Karpathy, L. F. Deep visual-semantic alignments for

generating image descriptions. Computer Vision and Pattern Recognition, Dec 2014.

[6] Hoo-Chang Shin, Kirk Roberts, L. L. D. D.-F. J. Y. R. M. S. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. Computer Vision and Pattern Recognition, Mar 2016.

[7] N Kausar, B. B. S., and Kuleev, R., 015. "Lung cancer detection using supervised classification with cluster variability on radiographs data". *ARPN Journal of Engineering and Applied Sciences,* **10**(20), Nov, p. 9274.

[8] Volodymyr Mnih, Nicolas Heess, A. G. k. k. Recurrent models of visual attention. NIPS, 2014.

[9] Jeff Donahue, Yangqing Jia, O. V. J. H. N. Z. E. T. T. D. Decaf: A deep convolutional activation feature for generic visual recognition. CVPR, Oct 2013.