Localizing tuberculosis in chest radiographs with deep learning

Zhiyun Xue[1], Stefan Jaeger[1], Sameer Antani[1], L. Rodney Long[1],
Alexandros Karargyris[2], Jenifer Siegelman[3], Les R. Folio[4]， George R. Thoma[1]

[1]National Library of Medicine, National Institutes of Health, Bethesda, MD
[2]IBM Research, Almaden, CA
[3]Harvard Medical School, Boston, MA and Takeda Pharmaceuticals, Cambridge, MA
[4]Clinical Center, National Institutes of Health, Bethesda, MD

ABSTRACT

Chest radiography (CXR) has been used as an effective tool for screening tuberculosis (TB). Because of the lack of radiological expertise in resource-constrained regions, automatic analysis of CXR is appealing as a "first reader". In addition to screening the CXR for disease, it is critical to highlight locations of the disease in abnormal CXRs. In this paper, we focus on the task of locating TB in CXRs which is more challenging due to the intrinsic difficulty of locating the abnormality. The method is based on applying a convolutional neural network (CNN) to classify the superpixels generated from the lung area. Specifically, it consists of four major components: lung ROI extraction, superpixel segmentation, multi-scale patch generation/labeling, and patch classification. The TB regions are located by identifying those superpixels whose corresponding patches are classified as abnormal by the CNN. The method is tested on a publicly available TB CXR dataset which contains 336 TB images showing various manifestations of TB. The TB regions in the images were marked by radiologists. To evaluate the method, the images are split into training, validation, and test sets with all the manifestations being represented in each set. The performance is evaluated at both the patch level and image level. The classification accuracy on the patch test set is 72.8% and the average Dice index for the test images is 0.67. The factors that may contribute to misclassification are discussed and directions for future work are addressed.

Keywords: Tuberculosis, Localization, Classification, Deep Learning

1. INTRODUCTION

Chest radiography (CXR) has been used as an effective tool for screening tuberculosis (TB) and various other pulmonary diseases. Early detection/treatment is key for reducing the spread of the disease. Because of the lack of radiological expertise in resource-constrained regions, automatic analysis of chest radiographs is appealing. There have been several efforts toward developing automatic screening systems for detecting TB infections using chest x-rays [1, 2]. At the U.S. National Library of Medicine (NLM), in collaboration with the AMPATH (Academic Model Proving Access to Healthcare), we have been developing a digital chest x-ray screening system for detecting manifestations consistent with exposure to TB in CXRs. The system is installed in a truck which serves parts of rural western Kenya. The general approach for CXR TB detection algorithms in the literature usually consists of three major steps: 1) Region-of-interest (ROI) identification: the ROI may be the whole image, the lung region, or a body rectangle that coarsely contains the lung; 2) Extracting features from the ROI: the features being used include texture features, shape features, and histogram-based features [2]; 3) Classification: a binary classifier (normal or abnormal) is then trained using these features. Typical classifiers used include support vector machines and random forest classifiers. Very recently, instead of using the general approach in which handcrafting features is a crucial step, a convolutional neural network (CNN) that learns the features automatically from the raw image data is used to classify TB CXR images [3]. Besides classifying a CXR to be normal or abnormal, it is also very important to identify location of TB in an abnormal CXR, as TB manifestations are often localized to a partial area of lung. In this paper, we focus on the task of locating/pinpointing TB in CXRs. We also explore a method based on CNN. Compared to the general field which has several large-scale labeled open image datasets, such as ImageNet and Microsoft COCO, one challenge of applying CNN to medical images is the lack of large annotated medical image datasets. One approach for taking advantage of the success of CNN in the general image domain is based on the idea of "transfer learning" [4]. In this approach, the CNN that has been pre-trained using a large scale, labeled, general image

domain dataset is used as a feature extractor for the images in the small, target dataset of interest. For our application, the number of annotated images is also quite small. However, we try to generate a big dataset of small image patches which can be used to train CNN directly. Specifically, we use patches that are extracted based on superpixels. In addition, we use multiscale patches in order to catch/preserve global spatial consistency. Therefore, the TB regions are located by identifying those superpixels whose corresponding patches are classified as abnormal by CNN. Work described in this manuscript is significantly different from [3] with respect to not only the task, but also the method although both approaches apply CNN for TB analysis in CXRs. In the following sections of the paper, we first introduce our dataset and the annotated collection. Then we describe our approach, followed by the presentation and discussion of our experimental tests. Finally, we conclude the paper with the outline of directions for future research.

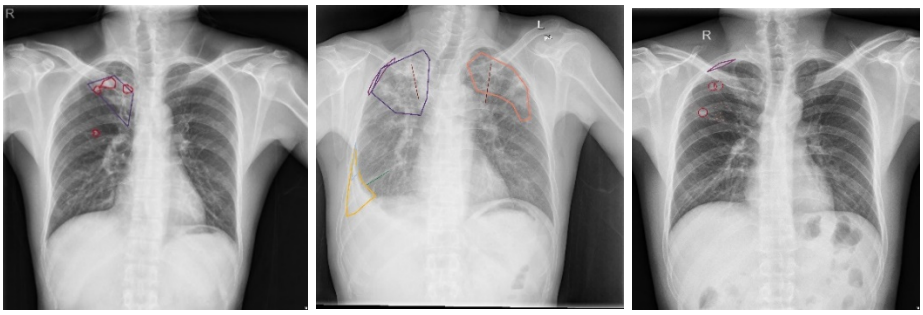


Fig. 1. Markings of abnormal areas by radiologists.

Table 1. Manifestations of TB in Shenzhen Dataset

Manifestation		Number of markings	Number of images
1.	Pleural effusion	45	41
2.	Apical thickening	126	98
3.	Single nodule (non-calcified)	170	60
4.	Pleural thickening	59	52
5.	Calcified nodule	237	79
6.	Small infiltrate (non-linear)	135	113
7.	Cavity	44	29
8.	Linear density	97	67
9.	Severe infiltrate (consolidation)	27	16
10.	Thickening of interlobar fissure	26	21
11.	Clustered nodule (2mm-5mm apart)	573	116
12.	Moderate infiltrate (non-linear)	47	35
13.	Calcification (other than nodule & lymph node)	13	6
14.	Calcified lymph node	9	6
15.	Miliary	4	3
16.	Retraction	29	21
17.	Adenopathy	14	9
18.	Other	8	6

2. DATASET

NLM has made two TB CXR datasets publicly available [5]. One is the Montgomery County CXR Set (MC) and the other is the Shenzhen Hospital CXR Set (Shenzhen). We worked on the Shenzhen dataset, since it contains many more images than the MC dataset. X-ray images in the Shenzhen data set have been collected as part of routine care by Shenzhen No.3

Hospital in Shenzhen, China. In this set, there are 326 normal X-rays and 336 abnormal TB X-rays showing various manifestations of TB. Image size varies for each X-ray, with approximately 3*K* by 3*K* pixels. In this study, we asked radiologists to mark the abnormal regions of all the TB X-rays in the Shenzhen set (336 images). There were two collaborating radiologists. Each was assigned half of the images. Each radiologist marked the pathology regions based on their observation on the image. The radiologists used the online tool Firefly [6] that was developed by the University of Missouri. Several examples of radiologists' markings are given in Figure 1. The radiologists used several ways to mark the abnormal sites depending on the manifestation types, such as b-spline curve, polygon, circle, line, and point. One image often contains several types of manifestations. The summarization of manifestations is given in Table 1. Specifically, in Table 1, the first column lists the 17 types of TB manifestations observed (plus "other") in the Shenzhen dataset; the second column lists the overall number of markings for each manifestation type; and the third column lists the number of images that contains that specific manifestation type. As shown in Table 1, the most frequently-occurring manifestations include nodules, small infiltrate, and apical thickening.

3. METHOD

Our previous work focuses on binary classification of an input chest radiograph as abnormal or normal [2]. The goal of our current work is to identify where the abnormal region is after an image is classified as abnormal. Our approach consists of four main steps: lung ROI extraction, superpixel segmentation, patch extraction and labeling, patch classification.

3.1 Lung ROI extraction

To reduce search space, the first step is to extract the ROI that includes the lungs. We have developed a non-rigid registration-driven lung segmentation algorithm. The method consists of three main stages: (i) a content-based image retrieval approach for identifying training images (with masks) most similar to the patient CXR using a partial Radon transform and Bhattacharyya shape similarity measure, (ii) creating the initial patient-specific anatomical atlas of lung shape using SIFT-flow for deformable registration of training masks to the patient CXR, and (iii) extracting refined lung boundaries using a graph cut optimization approach with a customized energy function. For the details of the lung segmentation algorithm, please refer to [7].

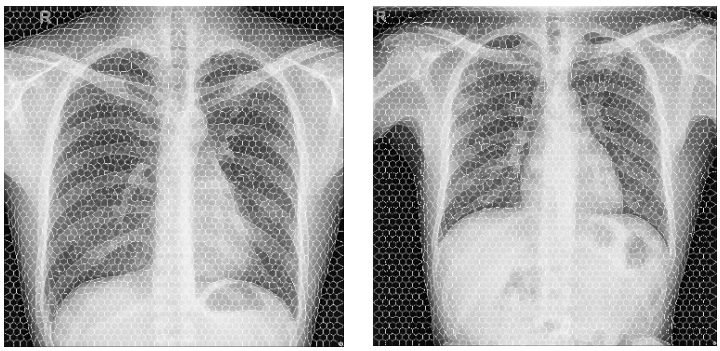


Fig. 2. Segmentation of superpixels.

3.2 Superpixel segmentation

To identify abnormal area, the lung ROI region is scanned to extract patches based on each superpixel instead of each pixel in the lung ROI region. The superpixels are generated using the Simple Linear Iterative Clustering (SLIC) method [8]. SLIC generates superpixels by clustering pixels based on their color similarity and proximity in the image plane. It controls the compactness and regularity of a superpixel using a distance measurement which provides balance between color similarity and spatial proximity. Two important parameters for SLIC are: the number of desired superpixels *k*, and the weighting factor between color and spatial differences *m*. Figure 2 shows two example results of superpixel segmentation.

3.3 Patch extraction and labeling

After the superpixels are generated, the next step is to extract patches. For each superpixel, its center is calculated and is used as the center of the patches to be extracted. At each center, three grayscale patches with size $L \times L$, $1.5L \times 1.5L$, and $2L \times 2L$ respectively are extracted. We wanted to capture image spatial consistency in a global sense, as well as capturing image details. This motivated us to use multiple patch sizes at each superpixel location. The three grayscale patches are combined to become a 3-channel color patch by resizing the patches with size $1.5L \times 1.5L$, and $2L \times 2L$ respectively to the size of $L \times L$. Generally speaking, the patch size ($L \times L$) should be larger than the superpixels and contain some context surrounding the superpixel. The color patch is labeled using the following rule. If any of the abnormal pixels marked by the radiologists (inside the marked polygon/circle region, or on the marked curve/line, or at the marked point), is inside the original L x L grayscale patch, then the patch is labeled as abnormal, otherwise the patch is labeled as normal.

3.4 Patch classification

We use a CNN to decide the abnormality of each superpixel. The input to the CNN is the three-channel color patch generated from a superpixel and the output of the CNN is the label of the color patch as defined in the previous paragraph. The well-known models of CNN include AlexNet, VGG, GoogLeNet, ResNet, etc. There are several open-source deep learning software. For example, Caffe, Tensorflow, Deeplearning4j, and Theano. In this paper, we use Caffe and adopt the CNN model used by Alex Krizhevsky for CIFAR-10 dataset [9] for our work. The CNN model is trained/validated/tested using patches extracted from the images in the training/validation/test set. The classified patches in a test image are then combined to generate the final binary mask showing the location of the abnormal regions.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the experimental test and results in detail. To reduce the computation time, the image is resized by ¼ in each dimension when applying SLIC superpixel segmentation. The parameter k is set to 3000 and m is set to 10. The generated superpixel mask is then resized back to the original image size (around $3K$ by $3K$). The average size of the superpixels is approximately 45×45. In the experiment, we set $L = 100$ for the patch size. For each image, three grayscale patches are extracted at each superpixel and a color patch is generated by combining these three grayscale patches. To train/test the CNN, we need to create the labeled training/validation/test patch datasets. To this end, we split the 336 abnormal TB images into three sets. We want to have images representing all the manifestations in each set. As shown in Table 1, the number of images for certain manifestations is very small. Specifically, there are 4 manifestations whose corresponding number of images is no larger than 6. They are: miliary, calcified lymph node, calcification (other than nodule & lymph node), and other. There are a total of 21 images exhibiting these manifestations. For each of the 4 manifestations, we randomly split its corresponding images into three sets. For example, for miliary, we put one of the 3 images in each set by random selection; for calcification (other than nodule & lymph node), there are a total of 6 images that exhibit this manifestation. We put 4 of the 6 images in the training set, one in the validation set and one in the test set by random selection. Note that no image exhibits more than one type of these 4 manifestations. Otherwise, we need to be careful to put each image into only one of the three sets so that the training/validation/test sets do not overlap. After splitting the images of these 4 manifestations, we then split the remaining 315 images randomly into training/validation/test sets based on a ratio of 70%/20%/10% approximately. As a result, there are 233, 67, and 36 images in the training, validation and test set, respectively. Table 2 lists the number of markings for each manifestation in each set. We then extract patches from the images in each of the three sets. We use the patches extracted from the superpixels within/overlapping with those abnormal markings as abnormal patches and the patches extracted from the superpixels that are inside the lung ROI but outside/not-overlapping with those abnormal markings as normal patches. Table 3 lists the number of abnormal/normal patches in each set. The number of abnormal patches is much smaller than that of normal patches. To obtain a balanced dataset, we did data augmentation on the abnormal patches as follows: we generated patches by shifting the center of each abnormal patch in 8 directions by 10 pixels and then randomly selected a certain equal number of patches in each direction so that the total number of the abnormal patches is about the same as that of normal patches in each set. We then use this augmented patch set to train and test the CNN model (which is the same as the model used for CIFAR-10 except that the input layer has dimension $100 \times 100 \times 3$ and the number of the outputs of the fully-connected layer is 2). The classification accuracy on the patch training and validation sets with the increasing of epoch is shown in Figure 3. The classification accuracy on the patch test set is 72.8% and the corresponding confusion matrix is given in Table 4. We also evaluate the performance on the image level. Specifically, for each test image, we generate a binary mask image based on the patch classification result (the superpixel is set as white/black if the corresponding patch is classified as abnormal/normal) and compare it with the ground truth mask by calculating the Dice index. The average Dice index for the test images is 0.67.

Figure 4 shows the results for one test image: (a) image with expert marking; (b) image with markings generated by the proposed method. The factors that may contribute to the misclassification include: (a) the intrinsic complexity/difficulty of identifying/locating the abnormality. For example, some abnormalities are subtle and hard to identify/locate; (b) the limited number of images, especially for certain manifestations; (c) the coarse delineations of abnormal areas by the experts for some cases. For example, to reduce labor intensity and the time involved, the radiologists marked polygons of some abnormal regions quite loosely and marked some nodules with only a point. Therefore, in the future, there are several directions that we could consider for improving the performance: using deeper and more complex models, obtaining more data and annotations, and refining the expert markings.

Table 2. Number of markings for each manifestation in each set

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
training	18	85	104	37	164	92	39	78	15
val.	20	31	51	16	40	28	4	13	7
test	7	10	15	6	33	15	1	6	5
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
training	12	399	30	4	7	1	17	8	6
val.	12	145	11	7	1	2	9	2	1
test	2	29	6	2	1	1	3	4	1

Table 3. Number of images/patches in each set

	Images	**Patches**	
		abnormal	***normal***
training	233	23,573	161,686
val.	67	9,859	41,709
test	36	4,409	23,897

Table 4. Confusion matrix for the patch test set

	Predicted	
	normal	***abnormal***
normal	17215	6682
abnormal	6304	17592

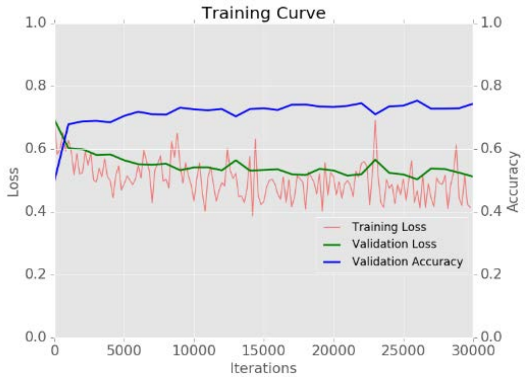
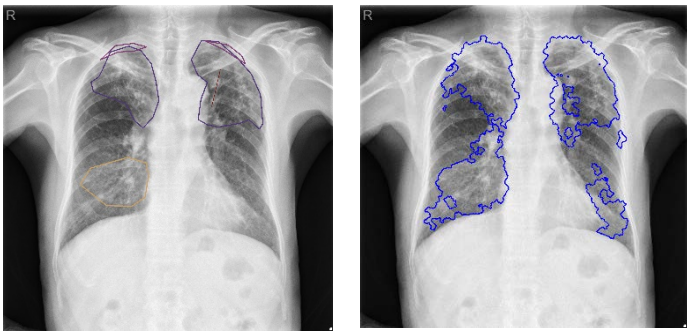


Fig. 3. Classification accuracy on the patch training and validation set.



(a) expert-marked image (b) result of the method

Fig. 4. Identified abnormal areas in the test image.

5. CONCLUSIONS

In this paper, we propose a method to address a very challenging task: to localize/pinpoint TB in a chest radiograph. The method is based on employing a CNN architecture to classify the superpixels generated from the lung area in the image. Specifically, it consists of four major components: lung ROI extraction, superpixel segmentation, patch generation and labeling, and patch classification. The method is tested on a publicly available dataset consisting of 336 TB images. The performance is evaluated at both the patch level and image level. The promising results demonstrate the effectiveness of the proposed method. Future work includes utilizing more complex/deeper CNN models and obtaining more annotated data for training.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

[1] http://www.diagnijmegen.nl/index.php/CAD4TB/.

[2] Jaeger, S., Karargyris, A., Candemir, S., Folio, L., et al., "Automatic tuberculosis screening using chest radiographs," IEEE Transactions on Medical Imaging 33 (2), 233-245 (2014).

[3] Lakhani, P., Sundaram, B., "Deep learning at chest radiography: automated classification of pulmonary Tuberculosis by using convolutional neural networks," Radiology 284(2), 574–582 (2017).

[4] Bar, Y., Diamant, I., Wolf, L., Greenspan, H., "Deep learning with non-medical training used for chest pathology identification," Proceedings of SPIE Medical Imaging, 94140V (2015).

[5] https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets/.

[6] http://firefly.cs.missouri.edu/

[7] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., et al., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," IEEE Transactions on Medical Imaging 33 (2), 577-590 (2014).

[8] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., S¨usstrunk, S., "SLIC superpixels – compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2274-2282 (2012).

[9] https://github.com/BVLC/caffe/tree/master/examples/cifar10/