Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN

Yuxi Dong[1], Yuchao Pan[1], Jun Zhang[2] and Wei Xu[1]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]Radiology department, The Fourth People's Hospital of Shaanxi, Shaanxi, China
summerdaway@gmail.com, panyuchao@gmail.com, weixu@tsinghua.edu.cn

***Abstract*—Chest radiography (chest X-ray) is a low-cost yet effective and widely used medical imaging procedures. The lacking of qualified radiologist seriously limits the applicability of the technique. We explore the possibility of designing a computer-aided diagnosis for chest X-rays using deep convolutional neural networks. Using a real-world dataset of 16,000 chest X-rays with natural language diagnosis reports, we can train a multi-class classification model from images and preform accurate diagnosis, without any prior domain knowledge.**

I. Introduction

Chest radiography (chest X-ray) is a medical imaging technology that is economical and easy to use. A chest X-ray produces an image of the chest, lung, heart, airways and blood vessels. Using a chest X-ray image, trained radiologist can diagnose conditions such as pneumonia, pneumothorax, interstitial lung disease, heart failure, bone fracture, hiatal hernia and so on.

The big advantage of X-ray is the low cost and simplicity to take. Photographing a chest X-ray is a non-invasive procedure that only takes a few minutes and the result typically comes out within half an hour. Modern digital radiography (DR) machines are quite affordable even in undeveloped regions. Thus, chest X-ray is widely used for screening as well as diagnostics of lung diseases such as lung nodules and interstitial lung diseases.

A large hospital typically produces over $40,000$ chest X-rays per year just from outpatient. Moreover, chest X-ray is a standard screening method in physical examinations that over three hundred million people took in 2014 all over China. This number is still increasing, resulting in hundreds of millions of chest X-ray images per year.

Lacking qualified radiologists to review these X-rays is a major challenge in China. Reviewing chest X-rays heavily depends on the experience of radiologists since the image has no spatial information and the overlap of different body parts may hide diseased tissues. Also, many images are difficult to read when the lesions are in low contrast or overlap with large pulmonary vessels. Even worse, each chest X-ray takes a trained radiologist several minutes to review and write the report, and many radiologists have to work over-time, increasing the misdiagnosis due to exhaustion.

As a result, the misdiagnosis of X-ray is high. It is reported that about 20% to 50% of lung nodules are missed or misdiagnosed on chest X-rays [1], while most of them can be detected retrospectively or by a second reviewer. Inexperienced radiologists are sometimes uncertain about their diagnosis, but they may not have the opportunity to discuss with others.

The recent development of artificial intelligence (AI), combining with the accumulating of a vast amount of medical images, opens a new opportunity to build an AI based system, aka. computer-aided diagnosis (CAD) system, to automatically analyze such chest X-rays. CAD covers many medical problems, such as automatic vertebrae detection [2], automatic coronary calcium scoring [3], lymph nodule (LN) detection [4], [5] and interstitial lung disease (ILD) classification [5], [6]. However, most recent CAD systems depends on the high resolution magnetic resonance imaging (MRI) or computed tomograph (CT) images, while ignoring the plain old X-rays that are much more widely utilized.

From image analysis point of view, CAD for X-rays is more challenging than MRIs or CTs. The layered images in MRIs and CTs reveal more details, and produce images with higher level of signal-to-noise ratio, making them easier to process. As a practical problem, while there are many X-ray images, there are very few X-ray datasets that are labeled as detailed as the MRI or CT datasets [7]–[10].

Traditional CADs are based on hand-crafted image features, and these features are then used to learn a binary or discrete classifier. The performance of such methods heavily depends on the extracted features, and it takes a long time for researchers to come up with a good set of features, especially for complicated images like the X-rays.

Deep convolutional neural network (CNN) has gained popularity given its excellent performance in different image recognition challenges, such as image classification [11]–[14] and semantic segmentation [15]–[18]. CNN is also applied in many medical image processing tasks [19]–[24] recently, and can reach the level of human radiologists [24].

CNN is a good fit for X-ray image analysis because it is an end-to-end network and easy to train. Also, CNN does not require any manual feature engineering - it only trains on the raw images and the classification labels (in our case, it is the diagnosis). Thus, as computer scientists without any radiology or medical background, we can build these models. Of course, the cost is the requirement of large amount of training data, which we can manage to get for X-ray images.

In our work, we use a dataset of $16,569$ chest X-ray images with their diagnosis reports to train a CNN model that can au-

tomatically generate these diagnoses on new images. Formally, this problem is called an image classification problem, where each disease is a class, and we want to classify each X-ray image into one or more classes.

To do so, we first automatically analyze the natural language report and extract the final diagnosis as *disease labels*. We allow multiple disease labels for each image. Then we train separate CNN models to perform three classification tasks: 1) is the image *normal*? 2) does this image has disease label X? and 3) what are all the disease labels of the image? For each task, we use two kinds of state-of-the-art CNNs and compare their accuracy results.

Our preliminary results shows that we can predict the disease labels with 82% accuracy in normal vs. abnormal task, and when we generate the top three most likely disease labels, we can predict the right label with over 97% accuracy. In multi-disease detection task, we achieve a mean average precision of 0.829.

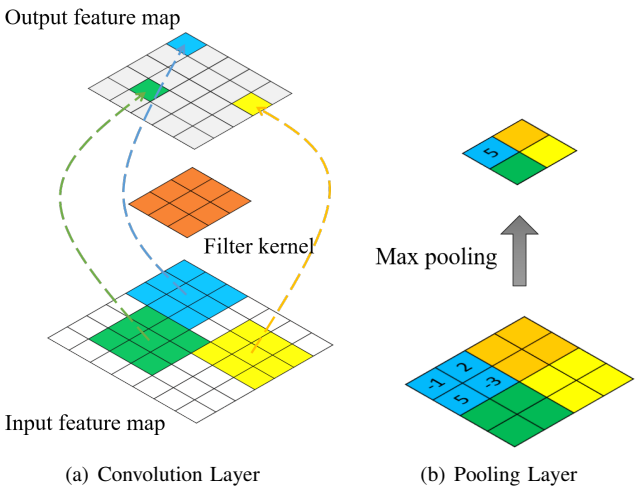
II. Related Work

An artificial neural network (ANN) approach was applied to the differential diagnosis of interstitial lung disease [25]. The artificial neural network was designed to distinguish between nine types of interstitial lung diseases based on 20 clinical items and radiographic information. The radiologists' performance in the differential diagnosis of interstitial lung disease was improved from an area under curve (AUC) of 0.826 to 0.911 in ROC curves by using the ANN output [26]. This work requires radiologists to read the X-ray image, and it takes a long time to extract the radiographic information. Our model is an end-to-end network, and the input of the network is the entire X-ray image without extra radiographic information.

Most work of CAD in medical imaging is based on feature extraction in a region-of-interest (ROI) in the image, such as histogram of oriented gradients (HoG [27]) and scale-invariant feature transform (SIFT [28]). These features are extracted directly based on RGB information, and are relatively low-level features and less expressive compared to convolutional neural networks (CNN) based features [29], [30].

Researchers apply CNNs on medical images in different imaging techniques recent years, such as magnetic resonance imaging (MRI) [19], [20], computed tomography (CT) [5], [21] and X-rays [22], [23]. Fine-tuning a pre-trained CNN model gains great success in many medical imaging problems. *H.C. Shin et al.* (2016) [5] fine-tune several ImageNet [31] pre-trained CNN models to study two specific CAD problems, thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification. *A. Esteva et al.* (2017) [24] also fine-tune an ImageNet pre-trained GoogLeNet [13] for skin cancer classification, reaching a level of competence comparable to dermatologists. We use a similar method to fine tune the pre-trained CNNs.

Previous work in [22] uses the off-the-shelf CNN (pre-trained CNN without fine-tuning) features combining with hand-crafted features to detect the chest pathology in X-rays.



(a) Convolution Layer (b) Pooling Layer

Fig. 1. The illustration of (a) Convolution Layer and (b) Pooling Layer. The output feature maps are computed by convolving filter kernels over the input feature maps. In pooling layer, every 2×2 patch is down-sampling by $\max$ operation.

They train an SVM of binary classification for each pathology and obtains an area under curve (AUC) of 0.87 for health vs pathology and 0.88-0.94 for different pathologies. *H.C. Shin et al.* (2016) [23] train RNNs to generate the annotations of chest X-rays, based on the fine-tuned CNN features. The fine-tuned CNNs achieve an accuracy of 66.40% on a 17-disease dataset. Our work does not require hand-crafted features either. Note that as none of these datasets are open, it is not possible to directly compare these accuracy results.

III. Background

Convolutional neural network (CNN) is a type of feed-forward neural network in machine learning. A convolutional neural network is formed by a stack of layers [11], [12], or a directed acyclic graph (DAG) of layers [13], [14]. A CNN usually combines the following five types of layers.

Convolution layers are the main components of a CNN. The layer consists of several filters (aka kernels) that we want to learn during the training face. Figure 1(a) shows the structure of a convolution layer. Formally, assuming $I = \{I_1, I_2, ..., I_N\}$ is the input feature maps and $K = \{K_1, K_2, ..., K_M\}$ where the size of K_i is $N \times W \times H$. The output feature maps $O = \{O_1, ..., O_M\}$ and $O_i = I \otimes K_i$ is computed as:

$$O_{i,x,y} = \sum_{j=1}^{N} \sum_{s=1}^{W} \sum_{t=1}^{H} I_{j,x+s,y+t} K_{i,j,s,t}$$

We usually add an *activation layer* after the convolution layer, increasing the non-linearity of the network. The *ReLu* (rectified linear unit) function $f(x) = \max(0, x)$ is used as a common practice by researchers [32]. Compared to other activation functions, such as *hyperbolic tangent* function $f(x) = \tanh(x)$ and *sigmoid* function $f(x) = (1 + e^{-x})^{-1}$, *ReLu* function, in spite of the hard non-linearity and non-

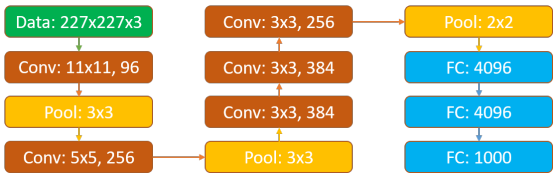


Fig. 2. The architecture of AlexNet. The network consists of 5 convolution layers, 3 pooling layers and 3 fully-connected layers. The parameters in convolution layers are the kernel size of the filters and the number of the output feature maps.

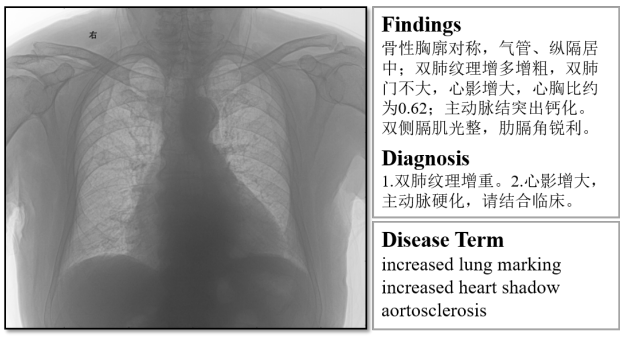


Fig. 3. An example of chest X-rays in our dataset. The report is written in Chinese and contains two parts of *findings* and *diagnosis*. The *disease labels* are extracted from *diagnosis*, and we show the result in English.

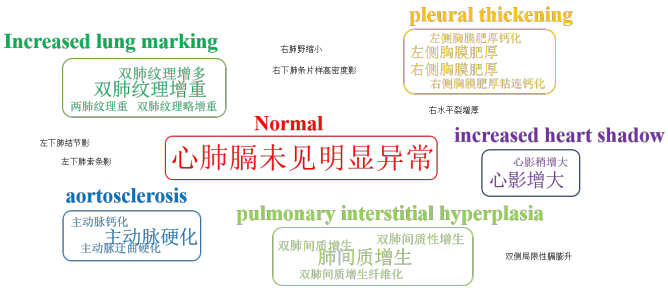


Fig. 4. Visualization of the clustering result. The font size is proportional to its appearance frequency. Boxes indicate different clusters.

differentiability at zero, creates sparse representations with true zeros and is suitable for naturally sparse data [33].

The *pooling layer* is another important concept of CNN, and it performs a non-linear down-sampling operation (Figure 1(b)). *Max-pooling* is the most commonly used pooling operation, and *Average-pooling* is also used according to the tasks.

The last layer of the network is usually a *fully-connected layer*. After several convolution and pooling layers, the network is ended by one or more fully-connected layers. Neurons in this layer have full connections with the previous layer. There is no spatial information after a fully-connected layer.

The *loss layer* is used to train the neural network. Various loss functions are used for different tasks. For example, *softmax loss* function is used for classification problem, and *sigmoid cross entropy loss* is used for predicting some independent probabilities.

In addition to the above layers, some effective techniques, such as batch normalization and dropout, are also used to improve the performance of CNNs. Figure 2 provides an example of CNN models called AlexNet. AlexNet takes a 227×227 RGB image as input, and produces a distribution over the 1000 class labels. We use an advanced versions of AlexNet, called VGG-Net [12], and another state-of-the-art CNN model ResNet [14].

IV. Methods

A. Dataset and pre-processing

We obtain a dataset with $16,569$ chest X-ray images taken on digital radiography (DR) machines in 2014 and 2015 from the fourth people's hospital of shaanxi in China. The hospital uses a modern Picture Archiving and Communication System (PACS) to store the images as well as their associated diagnosis reports. The images are in the Digital Imaging and Communications in Medicine (DICOM [34]) format. The diagnosis reports, aside from patient information (omitted for patient privacy in this study), contains two sections: the *findings* and *diagnosis*. All diagnosis reports are written in Chinese. Each diagnosis report is confirmed by a peer reviewer, which makes the report to be accurate. Figure 3 provides an example of the image with the report.

The *findings* sections in the reports describe the radiographic information, while *diagnosis* sections provide the radiologist-s' diagnoses of the diseases. These sections are in natural language. The hospital does not have a rigid coding system, and there may be several different terminologies to describe a single issue. There are also typos in these reports.

Thus, the first task is to convert these languages into accurate class labels. For pre-processing, we split the *diagnosis* into clauses, and we try to *cluster* similar clauses, which is highly likely to indicate similar diseases.

We define the similarity of the clauses based on their *edit distance* [35]. The *edit distance* is defined by the minimum number of operations (insertion, deletion, and substitution) to transform one clause into into the other. We find that in Chinese language, short edit distance is a pretty good indicator of semantically similar diagnosis. To compute the edit distance, we first remove the digits (i.e. variables) in sentences, and then use a dynamic programming algorithm to compute the distance.

Since it is hard to use a vector to represent a clause, we apply *k-medoids* algorithm [36] to perform the clustering on the clauses. *K-medoids* is related to the *k-means* [37] algorithm, and chooses the points in the dataset as the clustering centers. Figure 4 shows a visualization of the clustering result.

Given over 16 thousands of images, we have 10 diseases with 100 or more positive examples. Thus, we select the top ten largest clusters, covering 97% of all images, and assign a disease label for each, while ignoring the other smaller clusters. As we obtain more image samples, we will expand to more diseases as a future work. Table I shows the statistics of these clusters. The "single" column indicate that the label

TABLE I
TOP TEN MOST FREQUENT DISEASE LABELS

Disease Label	Total	Single	Single Percent
normal	8397	8397	100%
increased lung marking	6087	2927	48%
aortosclerosis	3930	743	19%
increased heart shadow	1159	88	8%
pleural thickening	630	120	19%
pulmonary interstitial hyperplasia	507	70	14%
costophrenic angle blunting	273	53	19%
pleural effusion	168	58	35%
emphysema	159	17	11%
bronchitis	144	77	53%

is the sole diagnosis of the image. For example, about 48% of the images (or $2,927$ images) are marked "increased lung marking" as the only disease, while the other 52% are also labeled with two or more diseases including "increased lung marking". Figure 5 provides an example of each normal and single disease images.

B. CNN-based classification model

Given CNN's layered structure, we can use a network to perform multiple tasks by changing its last layers. Specifically, we perform three tasks with the CNN: 1) classifying normal vs. abnormal images; 2) classifying images with a single disease; and 3) classifying multi-disease cases.

Although we are using over ten thousand images, the data size is still orders of magnitude smaller than normal image classification training data, such as the ImageNet. Thus, we use pre-trained models as a starting point, like many existing projects do [5], [24]. We obtain our pre-trained models on ImageNet from the Caffe Model Zoo [38]. Starting with these pre-trained models, we use the X-ray images as extra input, while training the model with the same set of training hyper-parameters such as the *batch size*, *learning rate*, and *momentum*.

We compare two state-of-the-art CNN models: VGG-16 (VGG-Net [12] with 16 layers) and ResNet-101 (residual network [14] with 101 layers). VGG-16 is a deeper version of AlexNet. It uses very small 3×3 size kernels in all convolution layers to reduce the number of parameters. ResNet-101 has a depth of up to 101 layers, much deeper than any other network. The key idea behind ResNet is addressing the vanishing gradient problem [39], [40] by computing the *residual* of a mapping instead of the mapping itself.

We rescale all images to a size of 256×256 since a classification CNN must take a fix-sized input because of the fully-connected layers.

Task 1: Classifying normal vs. abnormal images.

For this most basic task, we ignore the disease labels and consider all cases except for those marked as *normal* to be a single class. There are $8,397$ normal cases, and $8,172$ abnormal cases.

We randomly partition the dataset into training, validation and testing sets with $80\%/10\%/10\%$, and train the model on the training set, and use the resulting model to predict the binary classification result on the validation set.

Task 2: Multi class classification on images with single disease labels

As each image may indicate multiple diseases, we want the CNN model to predict the probability for each disease label. Thus, we modify the network, making the last layer of the CNN models to contain 10 neurons. Each neuron produces a probability distribution for a single disease label.

To reduce the noise when training the CNN for each disease, we use only the images with a single label as training and validation sets. We have $12,564$ such images (*Single* column of Table I). Except for the *emphysema* case, we have at least 50 images for each disease type. Thus we do have enough data to perform training, validation and testing for each label. Although we have 159 *emphysema* cases, they rarely present themselves as the single disease. Thus, we only use 5 cases for both validation and testing, while using all other 12 for training.

A big challenge here is the unbalanced datasets. For example 67.9% of single labels cases are *normal*, while we only have less than 1% of the images for diseases like *emphysema*. For model training, we need to balance the number of samples for each class, especially to boost the number of sample for the small classes. To do so, we adopt a common data augmentation technique in CNN [11]. We randomly crop 224×224 image patches from the 256×256 images and use each patch as a training sample. There are possibly $(256-224+1)^2 = 1,089$ patches for each image, and thus we can increase the number of samples for small classes. In our experiments, we boost the small classes so each class has a minimal of 380 samples, or 3% of all images. We also sample the normal cases, lowering its percentage in the training set to 21.9%.

Task 3: Classifying images with multiple disease labels

There are $3,879$ images with multiple disease labels. We find that the models trained on single disease cases are still useful in this case. We use the model to compute the probability of all labels, sort the probability in descending order and use all labels with probability above a threshold.

V. EXPERIMENTS

We perform our experiments on a Ubuntu server with 8 Titan X GPUs using the Caffe [41] framework. We perform all evaluation on the testing set (10%) of all the images, as described above. We present our evaluation results in this section.

A. Evaluation Metrics

We use the following metrics to evaluate the algorithm performance. First, we report the *sensitivity* and *specificity* of the detection, widely used in medical literature. We also introduce a metric called *top-3 sensitivity* following the widely used top-5 error ratio [42] in computer vision research. In *top-3 sensitivity* metric, a case is considered true positive if the true disease label is among the top three labels with the highest probability.

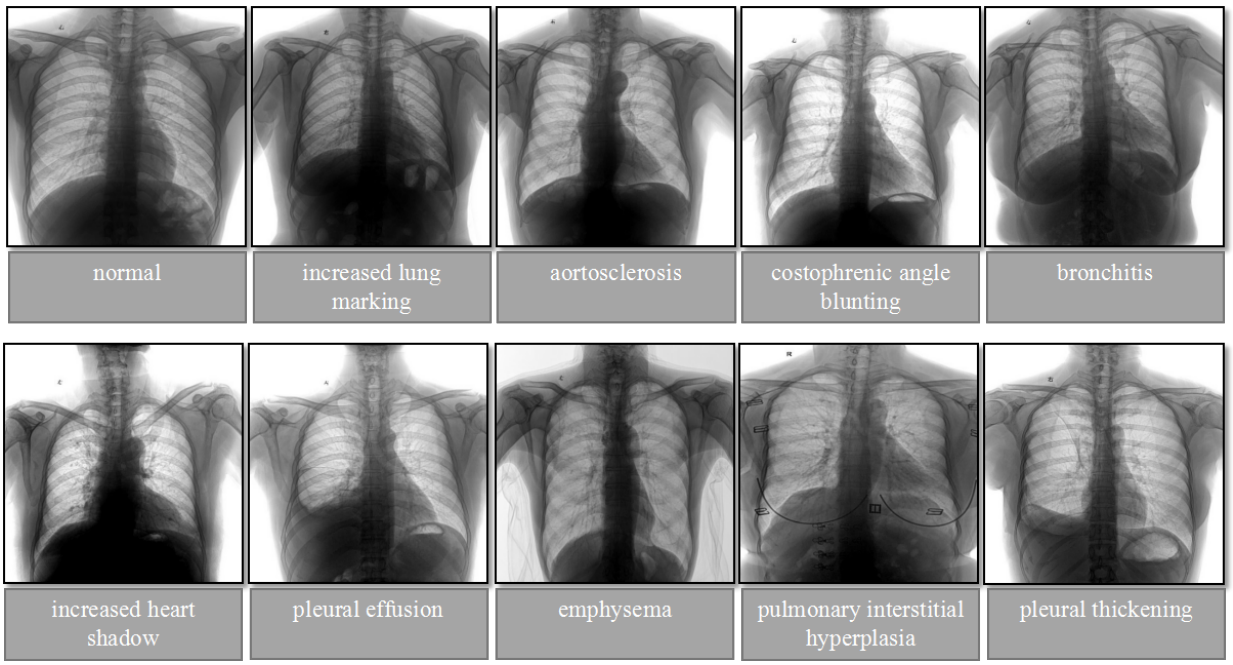
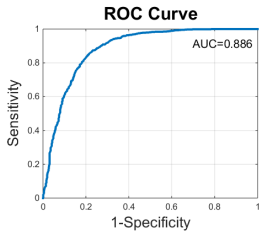


Fig. 5. Examples of chest X-rays for each disease.

Label	Predict	
	Normal	Abnormal
Normal	752	109
Abnormal	186	609

(a) Confusion matrix



(b) ROC curve

Fig. 6. The results of the confusion matrix and the ROC curve.

B. Task 1: Binary Classification on Normal vs. Abnormal

In this experiment, we use a testing set with $1,656$ images, out of which 52.0% are normal and 48.0% are *abnormal*. In this case, we only use the VGG-16 network. We train the network with 800 iterations, using 60 images per iteration with a training set size of $13,257$ images.

Figure 6(b) shows the confusion matrix and the ROC curve on the testing set. The accuracy in confusion matrix is 82.2%, and we achieve an AUC (i.e. area under curve) of 0.88 in the ROC curve.

C. Task 2: Single Disease Classification

During the test phase, following the method in AlexNet [11], the network makes the prediction by extracting five 224×224 patches, including the four corners and the center patch, and averaging the predictions on the five patches. Since we crop 224×224 patches for data augmentation in training set, the patch is more accurate than the entire image for the network. In the following experiments, the results are all obtained from the 5-patch testing method.

Table II summarizes the algorithm performance of each disease label. We can see that the algorithm performs well in some large cases, such as the normal and the aortosclerosis cases. ResNet generally performs better than VGG. The network performs well on most of the cases for the top-3 sensitivity cases. That is, the model is able to recognize the correct disease in the top three labels, reaching over 90% accuracy in many cases (where a random guess only gets about $3/10$ chance to be correct). As a CAD system, suggesting the three most likely labels can be very valuable in practice.

However, neither network performs well on the following cases, *increased heart shadow*, *pleural thickening*, *emphysema* and *bronchitis*. We believe the reasons are: 1) For *increased heart shadow*, our data augmentation method of cropping patches influences its detection, as the patches may not include the entire lung; 2) *Pleural thickening* usually appears with the symptom of *costophrenic angle blunting* in chest X-rays [43], and thus almost half of the cases of *pleural thickening* are misclassified into *costophrenic angle blunting*; and 3) *Bronchitis* is hard to diagnose from chest X-rays even for a human radiologist, as its diagnosis often depends on the full medical history [44].

D. Task 3: Multiple Disease Detection

In multi-disease detection cases, we use the images with at least two disease labels to test the models and see if the model can correctly predict all the disease labels.

For evaluation, we use the average precision (AP) [45] to evaluate the performance. The AP metric is based on the prediction list. Assume an X-ray image has m disease labels, and the rank for the disease labels in the prediction list are

TABLE II
SINGLE DISEASE CLASSIFICATION RESULTS

	VGG-16			ResNet-101		
Disease Label	Specificity	Sensitivity	Top-3 Sensitivity	Specificity	Sensitivity	Top-3 Sensitivity
normal	71.5%	78.5%	99.7%	70.5%	80.4%	99.5%
increased lung marking	79.4%	59.6%	100%	86.3%	43.2%	99.6%
aortosclerosis	97.0%	37.5%	92.4%	90.4%	61.1%	99.3%
increased heart shadow	99.8%	12.5%	50.0%	99.4%	25.0%	62.5%
pleural thickening	99.8%	16.7%	25.0%	99.6%	41.7%	58.3%
pulmonary interstitial hyperplasia	98.7%	57.1%	57.1%	99.9%	28.6%	42.9%
costophrenic angle blunting	99.9%	40.0%	60.0%	100%	60.0%	80.0%
pleural effusion	100%	60.0%	80.0%	99.9%	60.0%	100%
emphysema	99.8%	20.0%	20.0%	99.8%	20.0%	40.0%
bronchitis	99.6%	14.3%	28.6%	99.7%	28.6%	42.9%

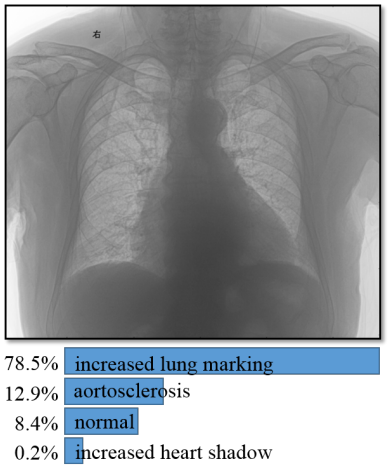


Fig. 7. An example of X-rays with multiple disease labels, and the disease labels could be found in Figure 3. The rank of the disease labels are 1, 2 and 4 in the list, so the average precision (AP) is $(1/1+2/2+3/4)/3 = 0.9167$.

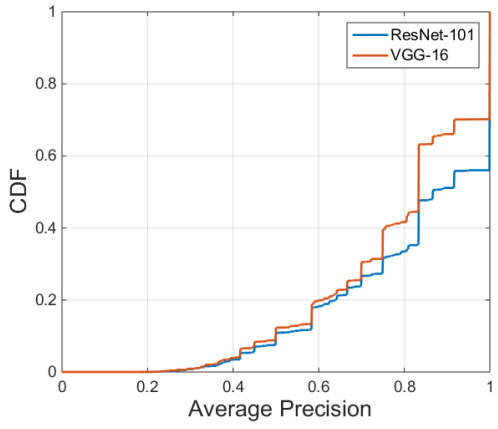


Fig. 8. The cumulative distribution functions.

$r_1, r_2, ..., r_m$, where $r_1 < ... < r_m$. Then the average precision (AP) of this case is computed as $\text{AP} = (1/m)\sum_i (i/r_i)$.

Figure 7 provides a concrete example of how the MAP metric is calculated. The human labels on the image are *increased lung marking*, *aortosclerosis* and *increased heart shadow* (the order does not matter). The model generates four labels with non-zero probability, as the figure shows. Then the AP for this image is calculated as $(1/1 + 2/2 + 3/4)/3 = 0.9167$.

Figure 8 presents a cumulative distribution function (CDF) of the AP metric across all images with at least two disease labels for both VGG and ResNet. As the previous experiments, ResNet performs better than VGG. The average AP value across these images is 0.829, meaning that the model can predict most of the disease labels in the dataset.

VI. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated our preliminary study on using CNN to train model for diagnosing diseases from chest X-ray images. We did this work from a purely computer science perspective: we train our model with no prior domain knowledge, but solely based on over 16,000 images with natural language diagnose reports. We have to deal with the highly unbalanced data problem using re-sampling method. We also fine-tune a pre-trained model to accelerate the training process. Using real world test dataset, we show that our method achieves very good accuracy.

As future work, we will use a larger dataset for training, especially adding positive examples of the relatively rare diseases. We will also incorporate the medical history into the model, better simulating how a human radiologist reads these images. Last but not least, we will use CNN to analyze multiple images types (X-ray, CT, MRI) from the same patient, linking their diagnostics.

VII. ACKNOWLEDGEMENT

This research is supported in part by the National Natural Science Foundation of China (NSFC) grant 61532001, Tsinghua Initiative Research Program Grant 20151080475, MOE Online Education Research Center (Quantong Fund) grant 2017ZD203, and gift funds from Huawei and Ant Financial.

REFERENCES

[1] J. V. Forrest and P. J. Friedman, "Radiologic errors in patients with lung cancer," *Western Journal of Medicine*, vol. 134, no. 6, pp. 485–490, 1981.

[2] H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Y. Cheng, and P. A. Heng, *Automatic Localization and Identification of Vertebrae in Spine CT via a Joint Learning Model with Deep Neural Networks*. Springer International Publishing, 2015.

[3] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks," *Medical Image Analysis*, vol. 9349, pp. 589–596, 2016.

[4] H. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation." *IEEE Transactions on Medical Imaging*, 2015.

[5] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 1285, 2016.

[6] S. Yang, W. Cai, H. Huang, Z. Yun, W. Yue, and D. D. Feng, "Locality-constrained subcluster representation ensemble for lung image classification," *Medical Image Analysis*, vol. 22, no. 1, pp. 102–113, 2015.

[7] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, and C. Ward, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging*, vol. 11, no. 7, pp. 757–771, 2008.

[8] A. S. Rd, G. Mclennan, M. F. Mcnittgray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, and H. Macmahon, "Lung image database consortium: developing a resource for the medical imaging research community." *Radiology*, vol. 232, no. 3, pp. 739–48, 2004.

[9] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Mller, "Building a reference multimedia database for interstitial lung diseases." *Computerized Medical Imaging & Graphics the Official Journal of the Computerized Medical Imaging Society*, vol. 36, no. 3, pp. 227–238, 2012.

[10] J. E. Burns, J. Yao, T. S. Wiese, H. E. Munoz, E. C. Jones, and R. M. Summers, "Automated detection of sclerotic metastases in the thoracolumbar spine at ct," *Radiology*, vol. 268, no. 1, pp. 69–78, 2013.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[13] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, "Going deeper with convolutions," pp. 1–9, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2015.

[15] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, p. 640, 2017.

[16] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3376–3385.

[17] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," pp. 1520–1528, 2015.

[18] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, no. 4, pp. 357–361, 2016.

[19] B. V. Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *IEEE International Symposium on Biomedical Imaging*, 2015, pp. 286–289.

[20] L. Rongjian, Z. Wenlu, S. Heung-Il, W. Li, L. Jiang, S. Dinggang, and J. Shuiwang, "Deep learning based imaging data completion for improved brain disease diagnosis," 2014, pp. 305–12.

[21] H. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation." *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1170–1181, 2016.

[22] Y. Bar, I. Diamant, L. Wolf, and S. Lieberman, "Chest pathology detection using deep learning with non-medical training," in *IEEE International Symposium on Biomedical Imaging*, 2015, pp. 294–297.

[23] H. C. Shin, K. Roberts, L. Lu, D. Demnerfushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2497–2506.

[24] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[25] N. Asada, K. Doi, H. Macmahon, S. M. Montner, M. L. Giger, C. Abe, and Y. Wu, "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: pilot study." *Radiology*, vol. 177, no. 3, pp. 857–60, 1990.

[26] S. Katsuragawa and K. Doi, "Computer-aided diagnosis in chest radiography," *Computerized Medical Imaging & Graphics*, vol. 31, no. 4-5, pp. 212–223, 2007.

[27] A. Seff, L. Lu, K. M. Cherry, H. R. Roth, J. Liu, S. Wang, J. Hoffman, E. B. Turkbey, and R. M. Summers, "2d view aggregation for lymph node detection using a shallow hierarchy of linear classifiers," in *Medical Image Computing & Computer-assisted Intervention: Miccai International Conference on Medical Image Computing & Computer-assisted Intervention*, 2014, pp. 544–552.

[28] M. Toews and T. Arbel, "A statistical parts-based model of anatomical variability," *IEEE Transactions on Medical Imaging*, vol. 26, no. 4, pp. 497–508, 2007.

[29] E. Simoserra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Morenonoguer, "Discriminative learning of deep convolutional feature point descriptors," in *IEEE International Conference on Computer Vision*, 2015, pp. 118–126.

[30] X. Han, T. Leung, Y. Jia, and R. Sukthankar, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[32] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. Lecun, "What is the best multi-stage architecture for object recognition?" in *Proc. International Conference on Computer Vision*, 2009, pp. 2146 – 2153.

[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Journal of Machine Learning Research*, vol. 15, 2011.

[34] C. Parisot, "The dicom standard," *International Journal of Cardiac Imaging*, vol. 11, no. 3, pp. 171–177, 1995.

[35] G. Navarro, "A guided tour to approximate string matching," *Acm Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2000.

[36] L. Kaufmann and P. J. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-norm & Related Methods*, 1987, pp. 405–416.

[37] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[38] "Caffe Model Zoo," https://github.com/BVLC/caffe/wiki/Model-Zoo, 2014.

[39] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[40] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets." *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[41] Y. Jia, E. Shelhamer, J. Donahue, and et al., "Caffe: Convolutional architecture for fast feature embedding," *Eprint Arxiv*, pp. 675–678, 2014.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[43] A. S. 3Rd, M. L. Giger, and H. Macmahon, "Computerized delineation and analysis of costophrenic angles in digital chest radiographs," *Academic Radiology*, vol. 5, no. 5, pp. 329–335, 1998.

[44] "How Is Bronchitis Diagnosed?" https://www.nhlbi.nih.gov/health/health-topics/topics/brnchi/diagnosis, 2011.

[45] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.