ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases

Xiaosong Wang[1], Yifan Peng [2], Le Lu [1], Zhiyong Lu [2], Mohammadhadi Bagheri [1], Ronald M. Summers [1]
[1]Department of Radiology and Imaging Sciences, Clinical Center,
[2] National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20892
Fxiaosong.wang,yifan.peng,le.lu,luzh,mohammad.bagheri,rmsG@nih.gov

Abstract

The chest X-ray is one of the most commonly accessible radiological examinations for screening and diagnosis of many lung diseases. A tremendous number of X-ray imaging studies accompanied by radiological reports are accumulated and stored in many modern hospitals' Picture Archiving and Communication Systems (PACS). On the other side, it is still an open question how this type of hospital-size knowledge database containing invaluable imaging informatics (i.e., loosely labeled) can be used to facilitate the data-hungry deep learning paradigms in building truly large-scale high precision computer-aided diagnosis (CAD) systems.

In this paper, we present a new chest X-ray database, namely "ChestX-ray8", which comprises 108,948 frontal-view X-ray images of 32,717 unique patients with the text-mined eight disease image labels (where each image can have multi-labels), from the associated radiological reports using natural language processing. Importantly, we demonstrate that these commonly occurring thoracic diseases can be detected and even spatially-located via a unified weakly-supervised multi-label image classification and disease localization framework, which is validated using our proposed dataset. Although the initial quantitative results are promising as reported, deep convolutional neural network based "reading chest X-rays" (i.e., recognizing and locating the common disease patterns trained with only image-level labels) remains a strenuous task for fully-automated high precision CAD systems.

1 Introduction

The rapid and tremendous progress has been evidenced in a range of computer vision problems via deep learning and large-scale annotated image datasets [25, 37, 13, 27]. Drastically improved quantitative performances in object recognition, detection and segmentation are demonstrated in comparison to previous shallow methodologies built upon hand-crafted image features. Deep neural network representations further make the joint language and vision learning tasks more feasible to solve, in image captioning [47, 23, 32, 46, 22], visual question answering [2, 45, 49, 53] and knowledge-guided transfer learning [4, 33], and so on. However, the intriguing and strongly observable performance gaps of the current state-of-the-art object detection and segmentation methods, evaluated between using PASCAL VOC [13] and employing Microsoft (MS) COCO [27], demonstrate that there is still significant room for performance improvement when underlying challenges (represented by different datasets) become greater. For example, MS COCO is composed of 80 object categories from 200k images, with 1.2M instances (350k are people) where every instance is segmented and many instances are small objects. Comparing to PASCAL VOC of only 20 classes and 11,530 images containing 27,450 annotated objects with bounding-boxes (BBox), the top competing object detection approaches achieve in 0.413 in MS COCO versus 0.884 in PASCAL VOC under mean Average Precision (mAP).

Figure 1. *Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.*

Deep learning yields similar rises in performance in the medical image analysis domain for object (often human anatomical or pathological structures in radiology imaging)

detection and segmentation tasks. Recent notable work includes (but do not limit to) an overview review on the future promise of deep learning [14] and a collection of important medical applications on lymph node and interstitial lung disease detection and classification [36, 42]; cerebral microbleed detection [11]; pulmonary nodule detection in CT images [39]; automated pancreas segmentation [35]; cell image segmentation and tracking [34], predicting spinal radiological scores [20] and extensions of multi-modal imaging segmentation [29, 16]. The main limitation is that all proposed methods are evaluated on some small-to-middle scale problems of (at most) several hundred patients. It remains unclear how well the current deep learning techniques will scale up to tens of thousands of patient studies.

In the era of deep learning in computer vision, research efforts on building various annotated image datasets [37, 13, 27, 2, 32, 53, 22, 24] with different characteristics play indispensably important roles on the better definition of the forthcoming problems, challenges and subsequently possible technological progresses. Particularly, here we focus on the relationship and joint learning of image (chest X-rays) and text (X-ray reports). The previous representative image caption generation work [47, 23] utilize Flickr8K, Flickr30K [51] and MS COCO [27] datasets that hold 8,000, 31,000 and 123,000 images respectively and every image is annotated by five sentences via Amazon Mechanical Turk (AMT). The text generally describes annotator's attention of objects and activity occurring on an image in a straightforward manner. Region-level ImageNet pre-trained convolutional neural networks (CNN) based detectors are used to parse an input image and output a list of attributes or "visually-grounded high-level concepts" (including objects, actions, scenes and so on) in [23, 49]. Visual question answering (VQA) requires more detailed parsing and complex reasoning on the image contents to answer the paired natural language questions. A new dataset containing 250k natural images, 760k questions and 10M text answers [2] is provided to address this new challenge. Additionally, databases such as "Flickr30k Entities" [32], "Visual7W" [53] and "Visual Genome" [24, 22] (as detailed as 94,000 images and 4,100,000 region-grounded captions) are introduced to construct and learn the spatially-dense and increasingly difficult semantic links between textual descriptions and image regions through the object-level grounding.

Though one could argue that the high-level analogy exists between image caption generation, visual question answering and imaging based disease diagnosis [41, 40], there are three factors making truly large-scale medical image based diagnosis (e.g., involving tens of thousands of patients) tremendously more formidable. **1**, Generic, open-ended image-level anatomy and pathology labels cannot be obtained through crowd-sourcing, such as AMT, which is prohibitively implausible for non-medically trained annotators. Therefore we exploit to mine the per-image (possibly multiple) common thoracic pathology labels from the image-attached chest X-ray radiological reports using Natural Language Processing (NLP) techniques. Radiologists tend to write more abstract and complex logical reasoning sentences than the plain describing texts in [51, 27]. **2**, The spatial dimensions of an chest X-ray are usually **2000 × 3000** pixels. Local pathological image regions can show hugely varying sizes or extents but often very small comparing to the full image scale. Fig. 1 shows eight illustrative examples and the actual pathological findings are often significantly smaller (thus harder to detect). Fully dense annotation of region-level bounding boxes (for grounding the pathological findings) would normally be needed in computer vision datasets [32, 53, 24] but may be completely nonviable for the time being. Consequently, we formulate and verify a weakly-supervised multi-label image classification and disease localization framework to address this difficulty. **3**, So far, all image captioning and VQA techniques in computer vision strongly depend on the ImageNet pre-trained deep CNN models which already perform very well in a large number of object classes and serves a good baseline for further model fine-tuning. However, this situation does not apply to the medical image diagnosis domain. Thus we have to learn the deep image recognition and localization models while constructing the weakly-labeled medical image database.

To tackle these issues, we propose a new chest X-ray database, namely "ChestX-ray8", which comprises 108,948 frontal-view X-ray images of 32,717 (collected from the year of 1992 to 2015) unique patients with the text-mined eight common disease labels, mined from the text radiological reports via NLP techniques. In particular, we demonstrate that these commonly occurred thoracic diseases can be detected and even spatially-located via a unified weakly-supervised multi-label image classification and disease localization formulation. Our initial quantitative results are promising. However developing fully-automated deep learning based "reading chest X-rays" systems is still an arduous journey to be exploited. Details of accessing the ChestX-ray8 dataset can be found in our website [1].

1.1 Related Work

There have been recent efforts on creating openly available annotated medical image databases [48, 50, 36, 35] with the studied patient numbers ranging from a few hundreds to two thousands. Particularly for chest X-rays, the largest public dataset is OpenI [1] that contains 3,955 radiology reports from the Indiana Network for Patient Care and 7,470 associated chest x-rays from the hospitals picture archiving and communication system (PACS). This database is utilized in [41] as a problem of caption generation but no quantitative disease detection results are reported. Our

[1] https://www.cc.nih.gov/drd/summers.html

newly proposed chest X-ray database is at least one order of magnitude larger than OpenI [1] (Refer to Table 1). To achieve the better clinical relevance, we focus to exploit the quantitative performance on weakly-supervised multi-label image classification and disease localization of common thoracic diseases, in analogy to the intermediate step of "detecting attributes" in [49] or "visual grounding" for [32, 53, 22].

2 Construction of Hospital-scale Chest X-ray Database

In this section, we describe the approach for building a hospital-scale chest X-ray image database, namely "ChestX-ray8", mined from our institute's PACS system. First, we short-list eight common thoracic pathology keywords that are frequently observed and diagnosed, i.e., Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumathorax (Fig. 1), based on radiologists' feedback. Given those 8 text keywords, we search the PACS system to pull out all the related radiological reports (together with images) as our target corpus. A variety of Natural Language Processing (NLP) techniques are adopted for detecting the pathology keywords and removal of negation and uncertainty. Each radiological report will be either linked with one or more keywords or marked with 'Normal' as the background category. As a result, the ChestX-ray8 database is composed of 108,948 frontal-view X-ray images (from 32,717 patients) and each image is labeled with one or multiple pathology keywords or "Normal" otherwise. Fig. 2 illustrates the correlation of the resulted keywords. It reveals some connections between different pathologies, which agree with radiologists' domain knowledge, e.g., Infiltration is often associated with Atelectasis and Effusion. To some extend, this is similar with understanding the interactions and relationships among objects or concepts in natural images [24].

2.1 Labeling Disease Names by Text Mining

Overall, our approach produces labels using the reports in two passes. In the first iteration, we detected all the disease concept in the corpus. The main body of each chest X-ray report is generally structured as "Comparison", "Indication", "Findings", and "Impression" sections. Here, we focus on detecting disease concepts in the Findings and Impression sections. If a report contains neither of these two sections, the full-length report will then be considered. In the second pass, we code the reports as "Normal" if they do not contain any diseases (not limited to 8 predefined pathologies).

Pathology Detection: We mine the radiology reports for disease concepts using two tools, DNorm [26] and MetaMap [3]. DNorm is a machine learning method for disease recognition and normalization. It maps every mention of keywords in a report to a unique concept ID in the Systematized Nomenclature of Medicine Clinical Terms (or SNOMED-CT), which is a standardized vocabulary of clinical terminology for the electronic exchange of clinical health information.

Figure 2. *The circular diagram shows the proportions of images with multi-labels in each of 8 pathology classes and the labels' co-occurrence statistics.*

MetaMap is another prominent tool to detect bio-concepts from the biomedical text corpus. Different from DNorm, it is an ontology-based approach for the detection of Unified Medical Language System® (UMLS®) Metathesaurus. In this work, we only consider the semantic types of Diseases or Syndromes and Findings (namely 'dsyn' and 'fndg' respectively). To maximize the recall of our automatic disease detection, we merge the results of DNorm and MetaMap. Table 1 (in the supplementary material) shows the corresponding SNOMED-CT concepts that are relevant to the eight target diseases (these mappings are developed by searching the disease names in the UMLS® terminology service [2], and verified by a board-certified radiologist.

Negation and Uncertainty: The disease detection algorithm locates every keyword mentioned in the radiology report no matter if it is truly present or negated. To eliminate the noisy labeling, we need to rule out those negated pathological statements and, more importantly, uncertain mentions of findings and diseases, e.g., "suggesting obstructive lung disease".

Although many text processing systems (such as [6]) can handle the negation/uncertainty detection problem, most of them exploit regular expressions on the text directly. One of the disadvantages to use regular expressions for negation/uncertainty detection is that they cannot capture vari-

[2] https://uts.nlm.nih.gov/metathesaurus.html

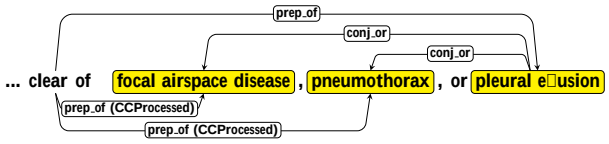

Figure 3. *The dependency graph of text: "clear of focal airspace disease, pneumothorax, or pleural effusion".*

ous syntactic constructions for multiple subjects. For example, in the phrase of "clear of A and B", the regular expression can capture "A" as a negation but not "B", particularly when both "A" and "B" are long and complex noun phrases ("clear of focal airspace disease, pneumothorax, or pleural effusion" in Fig. 3).

To overcome this complication, we hand-craft a number of novel rules of negation/uncertainty defined on the syntactic level in this work. More specifically, we utilize the syntactic dependency information because it is close to the semantic relationship between words and thus has become prevalent in biomedical text processing. We defined our rules on the dependency graph, by utilizing the dependency label and direction information between words.

As the first step of preprocessing, we split and tokenize the reports into sentences using NLTK [5]. Next we parse each sentence by the Bllip parser [7] using David McCloskys biomedical model [28]. The syntactic dependencies are then obtained from "CCProcessed" dependencies output by applying Stanford dependencies converter [8] on the parse tree. The "CCProcessed" representation propagates conjunct dependencies thus simplifies coordinations. As a result, we can use fewer rules to match more complex constructions. For an example as shown in Fig. 3, we could use "clear → prep_of → DISEASE" to detect three negations from the text □neg, focal airspace disease□, □neg, pneumothorax□, and □neg, pleural effusion□.

Furthermore, we label a radiology report as "normal" if it meets one of the following criteria:

- If there is no disease detected in the report. Note that here we not only consider 8 diseases of interest in this paper, but all diseases detected in the reports.

- If the report contains text-mined concepts of "normal" or "normal size" (CUIs C0205307 and C0332506 in the SNOMED-CT concepts respectively).

2.2 Quality Control on Disease Labeling

To validate our method, we perform the following experiments. Given the fact that no gold-standard labels exist for our dataset, we resort to some existing annotated corpora as an alternative. Using the OpenI API [1], we retrieve a total of 3,851 unique radiology reports where each OpenI report is assigned with its key findings/disease names by human annotators [9]. Given our focus on the eight diseases, a subset of OpenI reports and their human annotations are used as the gold standard for evaluating our method. Table 1 summarizes the statistics of the subset of OpenI [1, 19] reports. Table 2 shows the results of our method using OpenI, measured in precision (P), recall (R), and F1-score. Higher precision of 0.90, higher recall of 0.91, and higher F1-score of 0.90 are achieved compared to the existing MetaMap approach (with NegEx enabled). For all diseases, our method obtains higher precisions, particularly in "pneumothorax" (0.90 vs. 0.32) and "infiltration" (0.74 vs. 0.25). This indicates that the usage of negation and uncertainty detection on syntactic level successfully removes false positive imageupload. More importantly, the higher precisions meet our expectation to generate a Chest X-ray corpus with accurate semantic labels, to lay a solid foundation for the later processes.

Item #	OpenI	Ov.	ChestX-ray8	Ov.
Report	2,435	-	108,948	-
Annotations	2,435	-	-	-
Atelectasis	315	122	5,789	3,286
Cardiomegaly	345	100	1,010	475
Effusion	153	94	6,331	4,017
Infiltration	60	45	10,317	4,698
Mass	15	4	6,046	3,432
Nodule	106	18	1,971	1,041
Pneumonia	40	15	1,062	703
Pneumothorax	22	11	2,793	1,403
Normal	1,379	0	84,312	0

Table 1. *Total number (#) and # of Overlap (Ov.) of the corpus in both OpenI and ChestX-ray8 datasets.*

Disease	MetaMap			Our Method		
	P /	R /	F	P /	R /	F
Atelectasis	0.95 /	0.95 /	0.95	0.99 /	0.85 /	0.91
Cardiomegaly	0.99 /	0.83 /	0.90	1.00 /	0.79 /	0.88
Effusion	0.74 /	0.90 /	0.81	0.93 /	0.82 /	0.87
Infiltration	0.25 /	0.98 /	0.39	0.74 /	0.87 /	0.80
Mass	0.59 /	0.67 /	0.62	0.75 /	0.40 /	0.52
Nodule	0.95 /	0.65 /	0.77	0.96 /	0.62 /	0.75
Normal	0.93 /	0.90 /	0.91	0.87 /	0.99 /	0.93
Pneumonia	0.58 /	0.93 /	0.71	0.66 /	0.93 /	0.77
Pneumothorax	0.32 /	0.82 /	0.46	0.90 /	0.82 /	0.86
Total	0.84 /	0.88 /	0.86	0.90 /	0.91 /	0.90

Table 2. *Evaluation of image labeling results on OpenI dataset. Performance is reported using P, R, F1-score.*

2.3 Processing Chest X-ray Images

Comparing to the popular ImageNet classification problem, significantly smaller spatial extents of many diseases inside the typical X-ray image dimensions of **3000 × 2000** pixels impose challenges in both the capacity of computing hardware and the design of deep learning paradigm. In ChestX-ray8, X-rays images are directly extracted from the DICOM file and resized as **1024×1024** bitmap images without significantly losing the detail contents, compared with

image sizes of 512×512 in OpenI dataset. Their intensity ranges are rescaled using the default window settings stored in the DICOM header files.

2.4 Bounding Box for Pathologies

As part of the ChestX-ray8 database, a small number of images with pathology are provided with hand labeled bounding boxes (B-Boxes), which can be used as the ground truth to evaluate the disease localization performance. Furthermore, it could also be adopted for one/low-shot learning setup [15], in which only one or several samples are needed to initialize the learning and the system will then evolve by itself with more unlabeled data. We leave this as future work.

In our labeling process, we first select 200 instances for each pathology (1,600 instances total), consisting of 983 images. Given an image and a disease keyword, a board-certified radiologist identified only the corresponding disease instance in the image and labeled it with a B-Box. The B-Box is then outputted as an XML file. If one image contains multiple disease instances, each disease instance is labeled separately and stored into individual XML files. As an application of the proposed ChestX-ray8 database and benchmarking, we will demonstrate the detection and localization of thoracic diseases in the following.

3 Common Thoracic Disease Detection and Localization

Reading and diagnosing Chest X-ray images may be an entry-level task for radiologists but, in fact it is a complex reasoning problem which often requires careful observation and good knowledge of anatomical principles, physiology and pathology. Such factors increase the difficulty of developing a consistent and automated technique for reading chest X-ray images while simultaneously considering all common thoracic diseases.

As the main application of ChestX-ray8 dataset, we present a unified weakly-supervised multi-label image classification and pathology localization framework, which can detect the presence of multiple pathologies and subsequently generate bounding boxes around the corresponding pathologies. In details, we tailor Deep Convolutional Neural Network (DCNN) architectures for weakly-supervised object localization, by considering large image capacity, various multi-label CNN losses and different pooling strategies.

3.1 Unified DCNN Framework

Our goal is to first detect if one or multiple pathologies are presented in each X-ray image and later we can locate them using the activation and weights extracted from the network. We tackle this problem by training a multi-label DCNN classification model. Fig. 4 illustrates the DCNN architecture we adapted, with similarity to several previous weakly-supervised object localization methods [30, 52, 12, 18]. As shown in Fig. 4, we perform the network surgery on the pre-trained models (using ImageNet [10, 38]), e.g., AlexNet [25], GoogLeNet [44], VGGNet-16 [43] and ResNet-50 [17], by leaving out the fully-connected layers and the final classification layers. Instead we insert a transition layer, a global pooling layer, a prediction layer and a loss layer in the end (after the last convolutional layer). In a similar fashion as described in [52], a combination of deep activations from transition layer (a set of spatial image features) and the weights of prediction inner-product layer (trained feature weighting) can enable us to find the plausible spatial locations of diseases.

Multi-label Setup: There are several options of image-label representation and the choices of multi-label classification loss functions. Here, we define a 8-dimensional label vector $\mathbf{y} = [y_1, ..., y_c, ..., y_C], y_c \in \{0, 1\}, C = 8$ for each image. y_c indicates the presence with respect to according pathology in the image while a all-zero vector $[0, 0, 0, 0, 0, 0, 0, 0]$ represents the status of "Normal" (no pathology is found in the scope of any of 8 disease categories as listed). This definition transits the multi-label classification problem into a regression-like loss setting.

Transition Layer: Due to the large variety of pre-trained DCNN architectures we adopt, a transition layer is usually required to transform the activations from previous layers into a uniform dimension of output, $S \times S \times D, S \in \{8, 16, 32\}$. D represents the dimension of features at spatial location $(i, j), i, j \in \{1, ..., S\}$, which can be varied in different model settings, e.g., $D = 1024$ for GoogLeNet and $D = 2048$ for ResNet. The transition layer helps pass down the weights from pre-trained DCNN models in a standard form, which is critical for using this layers' activations to further generate the heatmap in pathology localization step.

Multi-label Classification Loss Layer: We first experiment 3 standard loss functions for the regression task instead of using the softmax loss for traditional multi-class classification model, i.e., Hinge Loss (HL), Euclidean Loss (EL) and Cross Entropy Loss (CEL). However, we find that the model has difficulty learning positive instances (images with pathologies) and the image labels are rather sparse, meaning there are extensively more '0's than '1's. This is due to our one-hot-like image labeling strategy and the unbalanced numbers of pathology and "Normal" classes. Therefore, we introduce the positive/negative balancing factor β_P, β_N to enforce the learning of positive examples. For example, the weighted CEL (W-CEL) is defined as follows,

$$L_{W\text{-}CEL}(f(\vec{x}), \vec{y}) = \beta_P \sum_{y_c=1} -\ln(f(x_c)) + \beta_N \sum_{y_c=0} -\ln(1 - f(x_c)), \quad (1)$$

where β_P is set to $\frac{|P|+|N|}{|P|}$ while β_N is set to $\frac{|P|+|N|}{|N|}$. $|P|$ and $|N|$ are the total number of '1's and '0's in a batch of image labels.

Figure 4. *The overall flow-chart of our unified DCNN framework and disease localization process.*

3.2 Weakly-Supervised Pathology Localization

Global Pooling Layer and Prediction Layer: In our multi-label image classification network, the global pooling and the predication layer are designed not only to be part of the DCNN for classification but also to generate the likelihood map of pathologies, namely a heatmap. The location with a peak in the heatmap generally corresponds to the presence of disease pattern with a high probability. The upper part of Fig. 4 demonstrates the process of producing this heatmap. By performing a global pooling after the transition layer, the weights learned in the prediction layer can function as the weights of spatial maps from the transition layer. Therefore, we can produce weighted spatial activation maps for each disease class (with a size of $S \times S \times C$) by multiplying the activation from transition layer (with a size of $S \times S \times D$) and the weights of prediction layer (with a size of $D \times C$).

The pooling layer plays an important role that chooses what information to be passed down. Besides the conventional max pooling and average pooling, we also utilize the Log-Sum-Exp (LSE) pooling proposed in [31]. The LSE pooled value x_p is defined as

$$x_p = \frac{1}{r} \cdot \log \left[\frac{1}{S} \cdot \sum_{(i,j) \in s} \exp(r \cdot x_{ij}) \right], \quad (2)$$

where x_{ij} is the activation value at (i, j), (i, j) is one location in the pooling region S, and $S = s \times s$ is the total number of locations in S. By controlling the hyper-parameter r, the pooled value ranges from the maximum in S (when $r \to \infty$) to average ($r \to 0$). It serves as an adjustable option between max pooling and average pooling. Since the LSE function suffers from overflow/underflow problems, the following equivalent is used while implementing the LSE pooling layer in our own DCNN architecture,

$$x_p = x^* + \frac{1}{r} \cdot \log \left[\frac{1}{S} \cdot \sum_{(i,j) \in s} \exp(r \cdot (x_{ij} - x^*)) \right], \quad (3)$$

where $x^* = \max\{|x_{ij}|, (i, j) \in S\}$.

Bounding Box Generation: The heatmap produced from our multi-label classification framework indicates the approximate spatial location of one particular thoracic disease class each time. Due to the simplicity of intensity distributions in these resulting heatmaps, applying an ad-hoc thresholding based B-Box generation method for this task is found to be sufficient. The intensities in heatmaps are first normalized to $[0, 255]$ and then thresholded by $\{60, 180\}$ individually. Finally, B-Boxes are generated to cover the isolated regions in the resulting binary maps.

4 Experiments

Data: We evaluate and validate the unified disease classification and localization framework using the proposed ChestX-ray8 database. In total, 108,948 frontal-view X-ray images are in the database, of which 24,636 images contain one or more pathologies. The remaining 84,312 images are normal imageupload. For the pathology classification and localization task, we randomly shuffled the entire dataset into three

subgroups for CNN fine-tuning via Stochastic Gradient Descent (SGD): i.e. training (70%), validation (10%) and testing (20%). We only report the 8 thoracic disease recognition performance on the testing set in our experiments. Furthermore, for the 983 images with 1,600 annotated B-Boxes of pathologies, these boxes are only used as the ground truth to evaluate the disease localization accuracy in testing (not for training purpose).

CNN Setting: Our multi-label CNN architecture is implemented using Caffe framework [21]. The ImageNet pre-trained models, i.e., AlexNet [25], GoogLeNet [44], VGGNet-16 [43] and ResNet-50 [17] are obtained from the Caffe model zoo. Our unified DCNN takes the weights from those models and only the transition layers and prediction layers are trained from scratch.

Due to the large image size and the limit of GPU memory, it is necessary to reduce the image *batch_size* to load the entire model and keep activations in GPU while we increase the *iter_size* to accumulate the gradients for more iterations. The combination of both may vary in different CNN models but we set $\texttt{batch\_size} \times \texttt{iter\_size} = 80$ as a constant. Furthermore, the total training iterations are customized for different CNN models to prevent over-fitting. More complex models like ResNet-50 actually take less iterations (e.g., 10000 iterations) to reach the convergence. The DCNN models are trained using a Dev-Box linux server with 4 Titan X GPUs.

Multi-label Disease Classification: Fig. 5 demonstrates the multi-label classification ROC curves on 8 pathology classes by initializing the DCNN framework with 4 different pre-trained models of AlexNet, GoogLeNet, VGG and ResNet-50. The corresponding Area-Under-Curve (AUC) values are given in Table 4. The quantitative performance varies greatly, in which the model based on ResNet-50 achieves the best results. The "Cardiomegaly" (AUC=0.8141) and "Pneumothorax" (AUC=0.7891) classes are consistently well-recognized compared to other groups while the detection ratios can be relatively lower for pathologies which contain small objects, e.g., "Mass" (AUC=0.5609) and "Nodule" classes. Mass is difficult to detect due to its huge within-class appearance variation. The lower performance on "Pneumonia" (AUC=0.6333) is probably because of lack of total instances in our patient population (less than 1% X-rays labeled as Pneumonia). This finding is consistent with the comparison on object detection performance, degrading from PASCAL VOC [13] to MS COCO [27] where many small annotated objects appear.

Next, we examine the influence of different pooling strategies when using ResNet-50 to initialize the DCNN framework. As discussed above, three types of pooling schemes are experimented: average looping, LSE pooling and max pooling. The hyper-parameter $\mathbf{r}$ in LSE pooling varies in $\{0.1, 0.5, 1, 5, 8, 10, 12\}$. As illustrated in Fig. 6, average pooling and max pooling achieve approximately equivalent performance in this classification task. The performance of LSE pooling start declining first when $\mathbf{r}$ starts increasing and reach the bottom when $\mathbf{r} = 5$. Then it reaches the overall best performance around $\mathbf{r} = 10$. LSE pooling behaves like a weighed pooling method or a transition scheme between average and max pooling under different $\mathbf{r}$ values. Overall, LSE pooling ($\mathbf{r} = 10$) reports the best performance (consistently higher than mean and max pooling).

Figure 5. *A comparison of multi-label classification performance with different model initializations.*

Figure 6. *A comparison of multi-label classification performance with different pooling strategies.*

Last, we demonstrate the performance improvement by using the positive/negative instances balanced loss functions (Eq. 1). As shown in Table 4, the weighted loss (W-CEL) provides better overall performance than CEL, especially for those classes with relative fewer positive instances, e.g. AUC for "Cardiomegaly" is increased from 0.7262 to 0.8141 and from 0.5164 to 0.6333 for "Pneumonia".

Disease Localization: Leveraging the fine-tuned DCNN

Setting	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
Initialization with different pre-trained models								
AlexNet	0.6458	0.6925	0.6642	0.6041	**0.5644**	0.6487	0.5493	0.7425
GoogLeNet	0.6307	0.7056	0.6876	0.6088	0.5363	0.5579	0.5990	0.7824
VGGNet-16	0.6281	0.7084	0.6502	0.5896	0.5103	0.6556	0.5100	0.7516
ResNet-50	**0.7069**	**0.8141**	**0.7362**	**0.6128**	0.5609	**0.7164**	**0.6333**	**0.7891**
Different multi-label loss functions								
CEL	0.7064	0.7262	0.7351	0.6084	0.5530	0.6545	0.5164	0.7665
W-CEL	0.7069	0.8141	0.7362	0.6128	0.5609	0.7164	0.6333	0.7891

Table 3. *AUCs of ROC curves for multi-label classification in different DCNN model setting.*

T(IoBB)	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
T(IoBB) = 0.1								
Acc.	0.7277	0.9931	0.7124	0.7886	0.4352	0.1645	0.7500	0.4591
AFP	0.0823	0.0487	0.0589	0.0426	0.0691	0.0630	0.0691	0.0264
T(IoBB) = 0.25 (Two times larger on both x and y axis than ground truth B-Boxes)								
Acc.	0.5500	0.9794	0.5424	0.5772	0.2823	0.0506	0.5583	0.3469
AFP	0.1666	0.1534	0.1189	0.0914	0.0975	0.0741	0.1250	0.0487
T(IoBB) = 0.5								
Acc.	0.2833	0.8767	0.3333	0.4227	0.1411	0.0126	0.3833	0.1836
AFP	0.2703	0.2611	0.1859	0.1422	0.1209	0.0772	0.1768	0.0772
T(IoBB) = 0.75								
Acc.	0.1666	0.7260	0.2418	0.3252	0.1176	0.0126	0.2583	0.1020
AFP	0.3048	0.3506	0.2113	0.1737	0.1310	0.0772	0.2184	0.0873
T(IoBB) = 0.9								
Acc.	0.1333	0.6849	0.2091	0.2520	0.0588	0.0126	0.2416	0.0816
AFP	0.3160	0.3983	0.2235	0.1910	0.1402	0.0772	0.2317	0.0904

Table 4. *Pathology localization accuracy and average false positive number for 8 disease classes.*

models for multi-label disease classification, we can calculate the disease heatmaps using the activations of the transition layer and the weights from the prediction layer, and even generate the B-Boxes for each pathology candidate. The computed bounding boxes are evaluated against the hand annotated ground truth (GT) boxes (included in ChestX-ray8). Although the total number of B-Box annotations (1,600 instances) is relatively small compared to the entire dataset, it may be still sufficient to get a reasonable estimate on how the proposed framework performs on the weakly-supervised disease localization task. To examine the accuracy of computerized B-Boxes versus the GT B-Boxes, two types of measurement are used, i.e, the standard Intersection over Union ratio (IoU) or the Intersection over the detected B-Box area ratio (IoBB) (similar to Area of Precision or Purity). Due to the relatively low spatial resolution of heatmaps ($\mathbf{32 \times 32}$) in contrast to the original image dimensions ($\mathbf{1024 \times 1024}$), the computed B-Boxes are often larger than the according GT B-Boxes. Therefore, we define a correct localization by requiring either $\mathbf{IoU > T(IoU)}$ or $\mathbf{IoBB > T(IoBB)}$. Refer to the supplementary material for localization performance under varying $\mathbf{T(IoU)}$. Table 4 illustrates the localization accuracy (Acc.) and Average False Positive (AFP) number for each disease type, with $\mathbf{T(IoBB) \in \{0.1, 0.25, 0.5, 0.75, 0.9\}}$. Please refer to the supplementary material for qualitative exemplary disease localization results for each of 8 pathology classes.

5 Conclusion

Constructing hospital-scale radiology image databases with computerized diagnostic performance benchmarks has not been addressed until this work. We attempt to build a "machine-human annotated" comprehensive chest X-ray database that presents the realistic clinical and methodological challenges of handling at least tens of thousands of patients (somewhat similar to "ImageNet" in natural images). We also conduct extensive quantitative performance benchmarking on eight common thoracic pathology classification and weakly-supervised localization using ChestX-ray8 database. The main goal is to initiate future efforts by promoting public datasets in this important domain. Building truly large-scale, fully-automated high precision medical diagnosis systems remains a strenuous task. ChestX-ray8 can enable the data-hungry deep neural network paradigms to create clinically meaningful applications, including common disease pattern mining, disease correlation analysis, automated radiological report generation, etc. For future work, ChestX-ray8 will be extended to cover more disease classes and integrated with other clinical information, e.g., follow-up studies across time and patient history.

Acknowledgements This work was supported by the Intramural Research Programs of the NIH Clinical Center and National Library of Medicine. We thank NVIDIA Corporation for the GPU donation.

References

[1] Open-i: An open access biomedical search engine. https://openi.nlm.nih.gov. 2, 3, 4

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, and L. Zitnick. Vqa: Visual question answering. In *ICCV*, 2015. 1, 2

[3] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, may 2010. 3

[4] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015. 1

[5] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. "O'Reilly Media, Inc.", 2009. 4

[6] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, oct 2001. 3

[7] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 173–180, 2005. 4

[8] M.-C. De Marneffe and C. D. Manning. *Stanford typed dependencies manual*. Stanford University, apr 2015. 4

[9] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, July 2015. 4

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5

[11] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. Mok, L. Shi, and P. Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE Trans. Medical Imaging*, 35(5):1182–1195, 2016. 2

[12] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. *IEEE CVPR*, 2016. 5

[13] M. Everingham, S. M. A. Eslami, L. J. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, pages 111(1): 98–136, 2015. 1, 2, 7

[14] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans. Medical Imaging*, 35(5):1153–1159, 2016. 2

[15] B. Hariharan and R. Girshick. Low-shot visual object recognition. *arXiv preprint arXiv:1606.02819*, 2016. 5

[16] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. Hemis: Hetero-modal image segmentation. In *MICCAI*, pages (2): 469–477. Springer, 2016. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 5, 7

[18] S. Hwang and H.-E. Kim. Self-transfer learning for weakly supervised lesion localization. In *MICCAI*, pages (2): 239–246, 2015. 5

[19] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wng, P.-X. Lu, and G. Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6), 2014. 4

[20] A. Jamaludin, T. Kadir, and A. Zisserman. Spinenet: Automatically pinpointing classification evidence in spinal mris. In *MICCAI*. Springer, 2016. 2

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 7

[22] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1, 2, 3

[23] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 2, 3

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 5, 7

[26] R. Leaman, R. Khare, and Z. Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015. 3

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and L. Zitnick. Microsoft coco: Common objects in context. *ECCV*, pages (5): 740–755, 2014. 1, 2, 7

[28] D. McClosky. *Any domain parsing: automatic domain adaptation for natural language parsing*. Thesis, Department of Computer Science, Brown University, 2009. 4

[29] P. Moeskops, J. Wolterink, B. van der Velden, K. Gilhuijs, T. Leiner, M. Viergever, and I. Isgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *MICCAI*. Springer, 2016. 2

[30] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE CVPR*, pages 685–694, 2015. 5

[31] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 6

[32] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1, 2, 3

[33] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, 2016. 1

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[35] H. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, pages 556–564. Springer, 2015. 2

[36] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In *MICCAI*, pages 520–527. Springer, 2014. 2

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 115(3): 211–252, 2015. 1, 2

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[39] A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. van Riel, M. Wille, M. Naqibullah, C. Snchez, and B. van Ginneken. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Trans. Medical Imaging*, 35(5):1160–1169, 2016. 2

[40] H. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers. Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *Journal of Machine Learning Research*, 17:1–31, 2016. 2

[41] H. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, 2016. 2

[42] H. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learnings. *IEEE Trans. Medical Imaging*, 35(5):1285–1298, 2016. 2

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 7

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 5, 7

[45] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *ICCV*, 2015. 1

[46] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *ICLR*, 2016. 1

[47] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1, 2

[48] H.-J. Wilke, M. Kmin, and J. Urban. Genodisc dataset: The benefits of multi-disciplinary research on intervertebral disc degeneration. In *European Spine Journal*, 2016. 2

[49] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 1, 2, 3

[50] J. Yao and et al. A multi-center milestone study of clinical vertebral ct segmentation. In *Computerized Medical Imaging and Graphics*, pages 49(4): 16–28, 2016. 2

[51] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014. 2

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *arXiv preprint arXiv:1512.04150*, 2015. 5

[53] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 1, 2, 3