IEEE Transactions on Medical Imaging

Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning

Journal:	*Transactions on Medical Imaging*
Manuscript ID	Draft
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Shin, Hoo-Chang; National Institutes of Health, Radiology and Imaging Sciences Roth, Holger; National Institutes of Health, Imaging Biomarkers and Computer-Aided Diagnosis Laboratory Gao, Mingchen; National Institutes of Health, Imaging Biomarkers and Computer-Aided Diagnosis Laboratory Lu, Le; National Institutes of Health, Radiology and Imaging Science Xu, Ziyue; National Institutes of Health, Department of Radiology and Imaging Sciences Yao, Jianhua; National Institute of Health, Radiology Department Mollura, Daniel J.; NIH, Summers, Ronald; National Institutes of Health Clinical Center, Radiology and Imaging Sciences;
Keywords:	Machine learning < General methodology, X-ray imaging and computed tomography < Imaging modalities, Abdomen < Object of interest, Lung < Object of interest, Computer-aided detection and diagnosis < General methodology, Evaluation and Performance < General methodology
Specialty/Area of Expertise:	machine learning, computer vision, medical imaging

Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning

Hoo-Chang Shin, *Member, IEEE*, Holger R. Roth, Mingchen Gao, Le Lu, *Senior Member, IEEE*, Ziyue Xu, Jianhua Yao, Daniel Mollura, Ronald M. Summers*

***Abstract*—Tremendous progress has been made in image recognition, primarily due to the availability of large-scale annotated datasets (i.e. ImageNet) and the rekindled development of deep convolutional neural networks (CNN). CNNs enable learning data-driven, highly representative, layered hierarchical image features given sufficient training data. However, obtaining datasets as comprehensively annotated as ImageNet in the medical imaging domain remains a challenge. There are currently two major techniques for employing the CNNs to medical image classification: training CNN from scratch by aggregating random views, or using off-the-shelf pre-trained CNN features. In this paper, we exploit three important, but previously understudied factors of employing deep convolutional neural networks to computer-aided detection problems.**

We first explore and evaluate different CNN architectures. The studied models contain 5 thousand to 160 million parameters, and vary in numbers of layers. We then evaluate the impacts on performance given datasets of different scales and spatial image contexts, and lastly, examine when and why transfer learning from pre-trained ImageNet CNNs (via fine-tuning) can be useful. Two specific computer-aided detection (CADe) problems, namely thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification, are studied. We achieve the state-of-the-art performance on the mediastinal LN detection, with 85% sensitivity at 3 false positive rate per patient, and report the first five-fold cross-validation classification results on predicting axial CT slices with ILD categories. Our extensive empirical evaluation, CNN model analysis and valuable insights can be generalized to design high performance CAD systems for other medical imaging tasks.

I. Introduction

Tremendous progress has been made in image recognition primarily due to the availability of large-scale annotated datasets (i.e. ImageNet [1], [12]) and the recently rekindled development of deep convolutional neural networks (CNN) [2], [3]. For data-driven learning, large-scale well-annotated datasets with representative data distribution characteristics are critical to learning more accurate or generalizable models [4], [3]. ImageNet [1] offers, for the first time in computer vision, a very comprehensive database of more than 1.2 million categorized natural images of 1000+ classes. The CNN models trained upon this database serve as the backbone for significantly improving many object detection and image segmentation problems using other datasets [5], [6], e.g., PASCAL [7] and medical image categorization [8], [9], [10], [11]. Yet, there exists no large-scale annotated medical image dataset comparable to the ImageNet, as data acquisition is difficult, and quality annotation is costly. There are currently two major techniques that successfully employ CNNs to medical image classification: *1)* training the "CNN from scratch" via decompositional 2D view resampling and aggregation of random view classification scores [13], [14]; or *2)* using "off-the-shelf CNN" features (without retraining the CNN) as the complementary information channels to existing hard-crafted image features, for Chest X-ray [9] and CT lung nodule identification [8], [11].

Despite the disparity between non-medical natural images and medical images, conventional image descriptors developed for object recognition in natural images such as scale-invariant feature transform (SIFT) [15] and histogram of oriented gradients (HOG) [16] have been widely used for object detection or segmentation in medical image analysis. Recently, ImageNet pre-trained CNNs have been used for chest pathology identification and detection in X-ray and CT modalities [9], [8], [11], achieving the best performed results by integrating low-level image features (e.g., GIST [17], bag of visual words (BoVW) and bag-of-frequency [11]). However fine-tuning a CNN model pre-trained on ImageNet upon specific medical image datasets has not been exploited or analyzed.

In this paper, we exploit three important, but previously under-studied factors of employing deep convolutional neural networks to computer-aided detection problems. Particularly, we explore and evaluate different CNN architectures ranging from 5 thousand to 160 million parameters with various of depths of CNN layers, describe the impacts of varying dataset scale and spatial image context on performance, and discuss when and why transfer learning from pre-trained ImageNet CNN models can be valuable. We further verify our hypothesis by inheriting or adapting rich hierarchical image features [4], [18] from the large-scale ImageNet dataset to computer aided diagnosis (CAD). We also explore CNN architectures of the most studied seven-layered "AlexNet-CNN" [3], a shallower "Cifar-CNN" [13], and a much deeper version of "GoogLeNet-CNN" [18] (with our modifications on CNN structures).

Two specific computer-aided detection (CADe) problems, namely thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification are studied in

Hoo-Chang Shin, Holger R. Roth, Le Lu, Jianhua Yao and Ronald M. Summers are with the Imaging Biomarkers and Computer-Aided Diagnosis Laboratory; Mingchen Gao, Ziyue Xu and Daniel Mollura are with Center for Infectious Disease Imaging, Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA. *Asterisk indicates corresponding author*. e-mail: {hoochang.shin, le.lu, rms}@nih.gov.

this work. On mediastinal LN detection, we achieve the best ever reported performance. We obtain 86% sensitivity on 3 false positives (FP) per patient, versus the prior state-of-art sensitivities of 78% [19] (stacked shallow learning) and 70% [13] (CNN), as prior state-of-the-art. For the first time, ILD classification results under the patient-level five-fold cross-validation protocol (CV5) is investigated and reported. The ILD dataset [20] contains 905 annotated image slices with 120 patients of 6 ILD labels. Such sparsely annotated datasets are generally difficult for CNN learning, due to the severe lack of labeled instances.

Evaluation protocols and details are critical to deriving significant empirical findings [21]. Our experimental results suggest that different CNN architectures and dataset re-sampling protocols are critical for the LN detection tasks (where labeled training data is relatively sufficient and spatial contexts are local), and that fine-tuning from ImageNet CNN models plays a minor role. However, fine-tuning fine-tuning ImageNet-trained models for ILD classification demonstrates clear advantage and early promising results where labeled training data is highly insufficient and multi-class categorization is used, as opposed to the LN dataset's binary class categorization. Another significant finding is that CNNs trained from scratch or fine-tuned from ImageNet models consistently outperform than using just off-the-shelf CNN features alone, in both the LN and ILD classification problems. We further analyze, via CNN activation visualizations, when and why transfer learning from non-medical to medical images in CADe problems can be valuable.

II. Datasets and Related Work

Using the three factors described above, we employ and evaluate CNNs under three aforementioned factors to thoracoabdominal lymph node (LN) detection (evaluated separately on mediastinal and abdominal regions) and interstitial lung disease (ILD) detection, in the image format of 2D CT slices [22], [23], [24] and randomly sampled 2.5D views in CT [13], [14], respectively. We then evaluate and compare CNN performance results.

Until the detection aggregation approach [13], [25], thoracoabdominal lymph node (LN) detection via CADe mechanisms has yielded poor performance results. In [13], each 3D LN candidates produces up to 100 random 2.5D orthogonally sampled images or views which are then used to train an effective CNN model. The best performance on abdominal LN detection is achieved at 83% recall on 3FP per patient [13], [14] using a "Cifar-10" CNN. Using the thoracoabdominal LN detection datasets [13], *we aim to surpass this CADe performance level, by testing different CNN architectures, exploring various dataset re-sampling protocols, and applying transfer learning from ImageNet pre-trained CNN models.*

Interstitial lung disease (ILD) refers to a group of more than 150 lung diseases affecting the interstitium, which can severely impair the patient's ability to breathe. Gao et al. [24] investigate the ILD classification problem in two scenarios: *1)* slice-level classification: assigning a holistic two-dimensional axial CT slice image with its occurring ILD disease label(s); and *2)* patch-level classification: first sampling patches within the 2D ROIs (Regions of Interest provided by [20]); followed by classifying patches into total seven category labels (i.e., six disease labels and healthy). Song et al. [22], [23] only address the second sub-task of patch-level classification under the "leave-one-patient-out" (LOO) criterion. In training upon the moderate-to-small scale ILD dataset [20], *our main objective is to exploit and benchmark CNN based ILD classification performances under the more realistic and unbiased CV5 metric (than LOO [22], [23] and hard-split [24]), with and without transfer learning.*

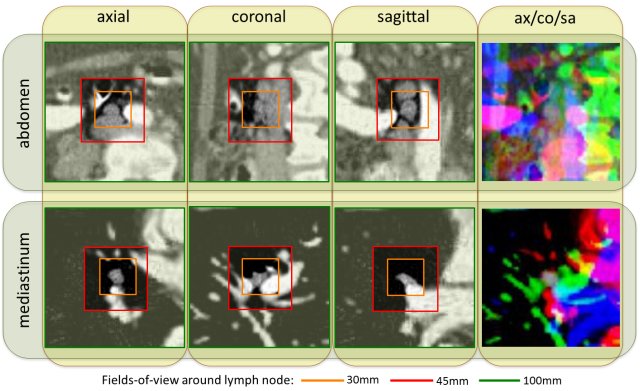
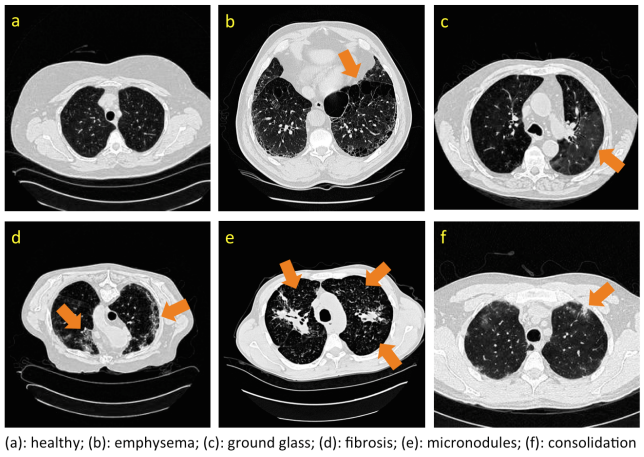


Fig. 1. Some examples of abdominal and mediastinal lymph nodes sampled on axial (ax), coronal (co), and sagittal (sa) views, with three different fields-of-views (30mm: orange; 45mm: red; 100mm: green) surrounding lymph nodes.

Thoracoabdominal Lymph Node Datasets. We use the publicly available dataset from [13], [25]. There are 388 mediastinal LNs labeled by radiologists in 90 patient CT scans, and 595 abdominal LNs in 86 patient CT scans. To facilitate comparison, we adopt the data preparation protocol of [13], where positive and negative LN candidates are sampled with the fields-of-view (FOVs) of 30mm to 45mm, surrounding the annotated/detected LN centers (obtained by a candidate generation process). In more details, [13], [25], [19] follow a coarse-to-fine CADe scheme, partially inspired by [26], which operates with $\sim$ 100% detection recalls at the cost of about 40 false or negative LN candidates per patient scan. In this work, positive and negative LN candidate are first sampled up to 200 times with translations and rotations. Afterwards, negative LN samples are randomly re-selected at a lower rate to roughly match the total number of positives. LN candidates are randomly extracted within the range of field-of-views (FOVs) of 35mm-128mm (so as to capture multiple spatial scales of image context [27]). They are then resized to a 64×64 pixel resolution with B-spline interpolation. Some examples of LNs with axial/coronal/sagittal views encoded in RGB color images [13] are shown in Figure 1.

Interstitial Lung Disease Dataset. We utilize the publicly available dataset of [20] of 120 patients and 905 image slices with six lung tissue types annotations with one or more of the following: healthy (NM), emphysema (EM), ground glass (GG), fibrosis (FB), micronodules (MN) and consolidation (CD) (Figure 3). At slice-level, the objective is to classify the status of "presence/absence" of six ILD classes for any



(a): healthy; (b): emphysema; (c): ground glass; (d): fibrosis; (e): micronodules; (f): consolidation

Fig. 2. Some examples of CT image slices with six lung tissue types in the ILD dataset [20]. Disease tissue types are located with dark orange arrows.

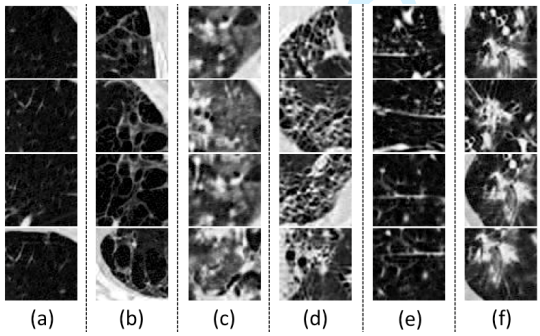


Fig. 3. Some examples of 64×64 pixel CT image patches for (a) NM, (b) EM, (c) GG, (d) FB, (e) MN (f) CD.

input axial CT slice [24]. Characterizing an arbitrary CT slice against any possible ILD types, without any manual ROI depicted (in contrast to [22], [23]), can be useful for large-scale patient screening. For slice-level ILD classification, we sampled the slices 12 times with random translations and rotations. After this, we balanced the numbers of CT slice samples for the six classes through random sampling of instances at various rates. For patch-based classification we sampled up to 100 patches of size 64×64 from each ROI. This dataset is divided into five folds with disjoint patient subsets. The average number of CT slices (training instances) per fold is small, as shown in Table I. Slice-level ILD classification is a very challenging task where CNN models need to learn from very small numbers of training examples and predict ILD labels on unseen patients.

To leverage the CNN architectures designed for color images and transfer CNN parameters pre-trained on ImageNet, we transform all gray-scale axial CT slice images via three CT windowing: [lung window range], [high-attenuation range], [low-attenuation range], and encode the transformed images into RGB channels. The low-attenuation CT windowing is useful to visualize certain texture patterns of lung diseases (especially emphysema). The high-attenuation range does not effectively highlight lung disease patterns so that we replace

normal	emphysema	ground glass	fibrosis	micronodules	consolidation
30.2	20.2	85.4	96.8	63.2	39.2

TABLE I
AVERAGE NUMBER OF IMAGES IN EACH FOLD FOR DISEASE CLASSES, WHEN DIVIDING THE DATASET IN 5-FOLD PATIENT SETS.

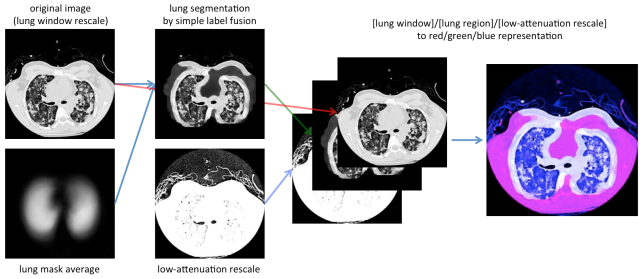


Fig. 4. Example of pre-processing the lung image slices. We convert the lung-window/lung-region/high-level-window settings to red/green/blue channels

it with a lung windowing performed inside the lung region.

Only rough lung segmentation is needed for this purpose that is achieved by segmenting the lung using simple label-fusion methods [28]. The procedure contains overlaying the average of lung mask in the training folds to the target image slice and performing simple morphology operations to get the lung boundary. Gaussian smoothing is then applied to the regions outside of the lung boundary to only preserve the details of inside lung. This process is illustrated in Figure 4.

III. METHODS

In this study, we fully exploit, evaluate and analyze the influence of various CNN Architectures, dataset characteristics (when we need more training data or better models for object detection [29]) and CNN transfer learning from non-medical to medical image domains. These three key elements of building effective deep CNN models for CADe problems are described below.

A. Convolutional Neural Network Architectures

We explore three convolutional neural network architectures (CifarNet [4], [13], AlexNet [3] and GoogLeNet [18]) with different model training parameter values. The current deep learning models [13], [30], [31] in medical image tasks are at least $2 \sim 5$ orders of magnitude smaller than even AlexNet [3]. More complex CNN models [13], [30] have only about 150K or 15K parameters. Roth et al. [13] adopts the CNN architecture tailored to the Cifar-10 dataset [4] and operates on image windows of $32 \times 32 \times 3$ pixels for lymph node detection, while the simplest CNN in [32] has only one convolutional/pooling/FC layer on $5 \times 5 \times 5$ or 32×32 pixel input windows, respectively.

We use CifarNet [4] as used in [13] as a baseline for the LN detection. AlexNet [3] and GoogLeNet [18] are also modified to evaluate these state-of-the-art CNN architecture from ImageNet classification task [12] to our CADe problems and datasets. A simplified illustration of three CNN architectures exploited is shown in Figure 6. CifarNet always takes $32 \times$

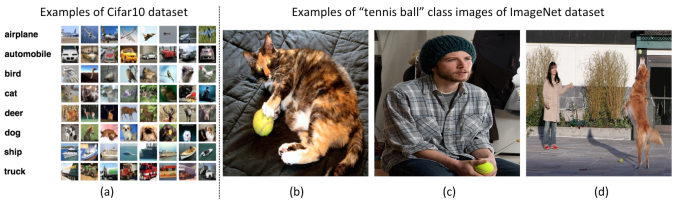


Fig. 5. Some examples of Cifar10 dataset and some images of "tennis ball" class from ImageNet dataset. Images of Cifar10 dataset are small (32×32) images with object of the image class category in the center. Images of ImageNet dataset are larger (256×256), where object of the image class category can be small, obscure, partial, and sometimes in a cluttered environment.

32×3 image patches as input while AlexNet and GoogLeNet are originally designed for the fixed image dimension of $256 \times 256 \times 3$ pixels. We also reduced the filter size, stride and pooling parameters of AlexNet and GoogLeNet to accomodate a smaller input size of $64 \times 64 \times 3$ pixels. We do so to produce and evaluate "simplified" AlexNet and GoogLeNet versions that are better suited to the smaller scale training datasets common in CADe problems. Throughout the paper, we refer to the models as CifarNet (32x32) or CifarNet (dropping 32x32); AlexNet (256x256) or AlexNet-H (high resolution); AlexNet (64x64) or AlexNet-L (low resolution); GoogLeNet (256x256) or GoogLeNet-H and GoogLeNet (64x64) or GoogLeNet-L (dropping 3 since all image inputs are three channels).

Additionally, to compare the method of using an ImageNet pre-trained AlexNet model on medical image analysis applications [9], [8], [11], we perform experiment on using "off-the-shelf" CNN features of AlexNet pre-trained on the ImageNet, and training only the final classifier layer to cope with new CADe classification tasks. Parameters in convolutional and fully connected layers are unchanged and used as deep image extractors, as in the same manner of [9], [8], [11]. We refer this model as AlexNet-ImNet in the remainder of the paper.

When learned from scratch, all the parameters of CNN models are initialized by random Gaussian distributions and trained for 30 epochs with the mini-batch size of 50 image instances. Training convergence can be observed within 30 epochs. The other hyperparameters are momentum: 0.9; weight decay: 0.0005; (base) learning rate: 0.01, decreased by a factor of 10 at every 10 epochs. We use the Caffe framework [33] and NVidia K40 GPUs to train the CNNs.

a) CifarNet: The CifarNet, introduced in [4], was the state-of-the-art model for object recognition on Cifar10 dataset, which consists of 32×32 images of 10 object classes. The objects are normally at the center ofthe images. Some example images and class categories from the Cifar10 dataset are shown in Figure 5. CifarNet has three convolution layers, three pooling layers, and one fully-connected layer. This CNN architecture architecture, also used in [13] has about 0.15 million free parameters. We adopt it as a baseline model for the LN detection.

b) AlexNet: The AlexNet architecture was published in [3], achieved significantly improved performance over the other non-deep learning methods for ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. This success

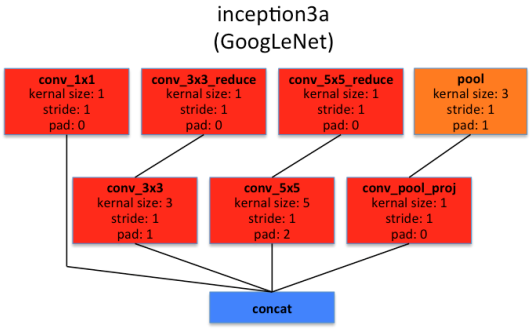


Fig. 7. Illustration of `inception3a` layer of GoogLeNet. Inception layers of GoogLeNet consist of six convolution layers with different kernel sizes and one pooling layer.

has rekindled interest in CNNs [2] in computer vision. ImageNet consists of 1.2 million 256×256 images belonging to 1000 categories. At times, the objects in the image are small and obscure, and thus pose more challenges for learning a successful classification model. More details about the ImageNet dataset will be discussed in Sec. III-B. AlexNet has five convolution layers, three pooling layers, and two fully-connected layers with about 60 million free parameters. This is our default CNN architecture for evaluation and analysis in the remainder of the paper.

c) GoogLeNet: The GoogLeNet model proposed in [18], is significantly more complex and deep than all previous CNN architectures. More importantly, it also introduces a new module called "inception", which concatenates filters of different sizes and dimension into a single new filter (refer to Figure 7). Overall, GoogLeNet has two convolution layers, two pooling layers, and nine inception layers. Each inception layer consists of six convolution layers and one pooling layer. An illustration of an inception layer (`inception3a`) from GoogLeNet is shown in Figure 7. GoogLeNet is the current state-of-the-art CNN architecture for the ILSVRC challenge, where it achieved 5.5% top-5 classification error on the ImageNet challenge, compared to AlexNet's 15.3% top-5 classification error.

B. ImageNet: Large Scale Annotated Natural Image Dataset

ImageNet [1] has more than 1.2 million 256×256 images which consists of 1000 object class categories. There are more than 1000 training images per class. The database is organized according to the WordNet [34] hierarchy, currently only nouns in thousand object categories. The image-object labels are obtained largely through crowd-sourcing, e.g., Amazon Mechanical Turk, and human inspection. Some examples of object categories in ImageNet are "sea snake", "sandwich", "vase", "leopard", etc. It is currently the largest image dataset among other standard datasets for visual recognition. Indeed, the Caltech101, Caltech256 and Cifar10 dataset merely contain 60000 32×32 images and 10 object classes. Furthermore, due to the large number (1000+) of object class, the objects belonging to each ImageNet class category can be occluded, partial and small, relative to those in the previous public image

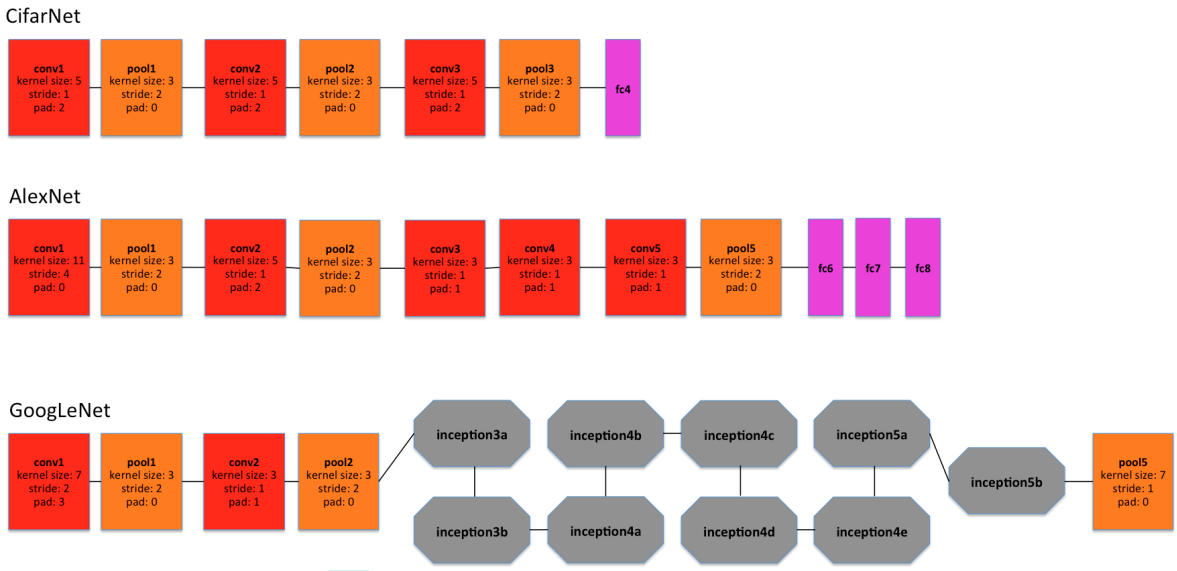


Fig. 6. A simplified illustration of the CNN architectures used. The inception layers of GoogLeNet [18] contains six convolution layers and one pooling layer. Each of the inception layer of GoogLeNet consists of six convolution layers and one pooling layer.

datasets. This significant intra-class variation poses greater challenge to any data-driven learning system that builds a classifier to fit given data and generalize to unseen data. As a comparison, some example images of Cifar10 dataset and ImageNet images in the "tennis ball" class category are shown in Figure 5. The ImageNet dataset is publicly available, and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has become the standard benchmark for large-scale object recognition.

C. Transfer Learning

AlexNet and GoogLeNet CNN models can be either learned from scratch or fine-tuned from pre-trained models. Girshick et al. [5] find that, by applying ImageNet pre-trained ALexNet to PASCAL dataset [7], performances of semantic 20-class object detection and segmentation tasks significantly improve over previous methods that use no deep CNNs. AlexNet can be fine-tuned on the PASCAL dataset to surpass the performance of the ImageNet pre-trained AlexNet, although the difference is not as significant as that between the CNN and non-CNN methods. Similarly, [35], [36] also demonstrate that better performing deep models are learned via CNN transfer learning from ImageNet to other datasets of limited scales.

Our hypothesis on CNN parameter transfer learning is the following: despite the disparity between natural images and natural images, CNNs comprehensively trained on the large scale well-annotated ImageNet may still be transferred to make medical image recognition tasks more effective. Collecting and annotating large numbers of medical images still poses significant challenges. On the other hand, the mainstream deep CNN architectures (e.g., AlexNet and GoogLeNet) contain tens of millions of free parameters to train, and thus require sufficiently large numbers of labeled medical images.

For transfer learning, we follow the approach of [35], [5] where all CNNs layers except the last are fine-tuned at a learning rate 10 times smaller than the default learning rate. The last fully-connected layer is random initialized and freshly trained, In order to accommodate the new object categories in our CADe applications. Its learning rate is kept at the original 0.01. We denote the models with the random initialization or transfer learning as AlexNet-RI, AlexNet-TL, GoogLeNet-RI and GoogLeNet-TL.

IV. Evaluations and Discussions

In this section, we evaluate and compare the performances of nine configurations of CNN models (CifarNet, AlexNet-ImNet, AlexNet-RI-H, AlexNet-TL-H, AlexNet-RI-L, GoogLeNet-RI-H, GoogLeNet-TL-H, GoogLeNet-RI-L and combined) on two important CADe problems using publicly available datasets [13], [25], [20].

A. Thoracoabdominal Lymph Node Detection

We train and evaluate CNNs using three-fold cross-validation (folds are split into disjoint sets of patients), with the different CNN architectures described above. In testing, each LN candidate has multiple random 2.5D views tested by CNN classifiers to generate LN class probability scores. We follow the random view aggregation through averaging probabilities, as in [13].

We first sample the LN image patches at a 64×64 pixel resolution. We then up-sample the 64×64 pixel LN images via bi-linear interpolation to 256×256 pixels, in order to accommodate AlexNet-RI-H, AlexNet-TL-H, GoogLeNet-RI-H and GoogLeNet-TL-H. For the modified AlexNet-RI-L at (64×64) pixel resolution, we reduce the number of first layer convolution filters from 96 to 64 and also reduce the stride from 4 to 2. For modified GoogLeNet-RI (64×64), we drop the number of first layer convolution filters from 64 to 32, pad size from 3 to 2, kernel size from 7 to 5, stride from 2 to 1

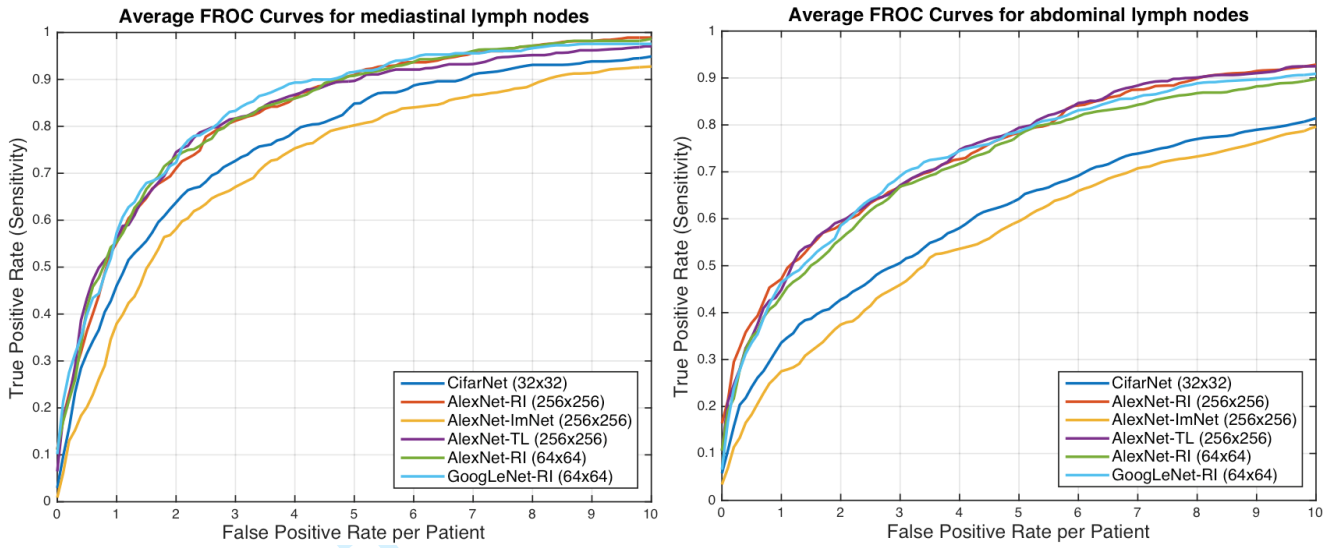


Fig. 8. FROC curves averaged on three-fold CV for the abdominal (left) and mediastinal (right) lymph nodes using different CNN models.

Region	Mediastinum		Abdomen	
Method	AUC	TPR/3FP	AUC	TPR/3FP
[25]	-	0.63	-	0.70
[13]	**0.92**	0.70	**0.94**	**0.83**
[19]	-	**0.78**	-	0.78
CifarNet	0.91	0.70	0.81	0.44
AlexNet-ImNet	0.89	0.63	0.80	0.41
AlexNet-RI-H	0.94	0.79	0.92	0.67
AlexNet-TL-H	0.94	0.81	0.92	0.69
GoogLeNet-RI-H	0.85	0.61	0.80	0.48
GoogLeNet-TL-H	0.94	0.81	**0.92**	**0.70**
AlexNet-RI-L	0.94	0.77	0.88	0.61
GoogLeNet-RI-L	**0.95**	**0.85**	0.91	0.69
Combined	0.95	0.85	0.93	0.70

TABLE II
COMPARISON OF MEDIASTINAL AND ABDOMINAL LN DETECTION RESULTS USING VARIOUS CNN MODELS. BOLD NUMBERS INDICATE THE BEST PERFORMANCE VALUES ON CLASSIFICATION ACCURACY.

and the stride of the subsequent pooling layer from 2 to 1. CifarNet is used in [13] to detect LN samples of $32 \times 32 \times 3$ images. For consistency reasons, we down-sample $64 \times 64 \times 3$ resolution LN sample images to the dimension of $32 \times 32 \times 3$.

Results for lymph node detection in the mediastinum and abdomen are reported in Table II. FROC curves are illustrated in Figure 8. The area-under-the-FROC-curve (AUC) and true positive rate (TPR, recall or sensitivity) at three false positives per patient (TPR/3FP) are used as performance metrics. Out of nine investigated CNN models, CifarNet, AlexNet-ImNet and GoogLeNet-RI-H in general yielded the least competitive detection accuracy. Our LN datasets are significantly more complex (i.e., display much larger within-class appearance variations), especially with the extracted fields-of-view (FOVs) of (35mm-128mm) compared to (30mm-45mm) in [13], where CifarNet is also employed. In this experiment, CifarNet is under-trained with respect to our enhanced LN datasets, due to its limited input resolution and parameter complexity. The inferior performance of AlexNet-ImNet implies that using the pre-trained ImageNet CNNs alone as "off-the-shelf" deep image feature extractors may not be optimal, or not adequate for mediastinal and abdominal LN detection tasks. To compensate "off-the-shelf" CNN features, [9], [8], [11] all add and integrate various other hand-crafted image features as hybrid inputs for the final CADe classification.

GoogLeNet-RI-H performs poorly, as it is susceptible to over-fitting. No sufficient data samples are available to train GoogLeNet-RI-H from random initialization. Indeed, due to GoogLeNet-RI-H's complexity and 22-layer depth, million-image datasets may be required to properly train this model. However GoogLeNet-TL-H significantly improves upon GoogLeNet-RI-H (0.81 versus 0.61 TPR/3FP in mediastinum; 0.70 versus 0.48 TPR/3FP in abdomen), thus indicating that transfer learning offers a much better initialization of CNN parameters than random initialization. Likewise, AlexNet-TL-H consistently outperforms AlexNet-RI-H but at smaller margins (0.81 versus 0.79 TPR/3FP in mediastinum; 0.69 versus 0.67 TPR/3FP in abdomen). This is also consistent with the findings reported for ILD detection in Table III and Figure 9.

GoogLeNet-TL-H yields results similar to AlexNet-TL-H's for in the mediastinal LN detection, but slightly outperforms Alex-Net-H for abdominal LN detection. AlexNet-RI-H shows less severe over-fitting than GoogLeNet-RI-H. We also evaluate a simple ensemble by averaging the probability scores from five CNNs: AlexNet-RI-H, AlexNet-TL-H, AlexNet-RI-H, GoogLeNet-TL-H and GoogLeNet-RI-L. This combined ensemble outputs the classification accuracies matching or slightly exceeding the best performing individual CNN models on the mediastinal or abdominal LN detection tasks, respectively.

Many of our CNN models achieve notably better (FROC-AUC and TPR/3FP) results than the previous state-of-the-art models [19] for **mediastinal** LN detection: GoogLeNet-RI-L obtains an AUC=0.95 and 0.85 TPR/3FP, versus AUC=0.92

	NM	EM	GG	FB	MN	CD
Patch-LOO [22]	0.84	0.75	0.78	0.84	0.86	-
Patch-LOO [23]	0.88	0.77	0.80	0.87	0.89	-
Patch-CV10 [32]	0.84	0.55	0.72	0.76	0.91	-
Patch-CV5	0.64	0.81	0.74	0.78	0.82	0.64
Slice-Test [24]	0.40	1.00	0.75	0.80	0.56	0.50
Slice-CV5	0.22	0.35	0.56	0.75	0.71	0.16
Slice-Random	0.90	0.86	0.85	0.94	0.98	0.83

TABLE IV
COMPARISON OF INTERSTITIAL LUNG DISEASE CLASSIFICATION RESULTS USING F-SCORES: NM, EM, GG, FB, MN AND CD.

Ground truth	Prediction					
	NM	EM	GG	FB	MN	CD
NM	**0.68**	0.18	0.10	0.01	0.03	0.01
EM	0.03	**0.91**	0.00	0.02	0.03	0.01
GG	0.06	0.01	**0.70**	0.09	0.06	0.08
FB	0.01	0.02	0.05	**0.83**	0.05	0.05
MN	0.09	0.00	0.07	0.04	**0.79**	0.00
CD	0.02	0.01	0.10	0.18	0.01	**0.68**

TABLE V
CONFUSION MATRIX FOR ILD CLASSIFICATION (PATCH-LEVEL) WITH FIVE-FOLD CV USING GOOGLENET-TL.

and 0.70 TPR/3FP [13] and 0.78 TPR/3FP [19] which uses stacked shallow learning. This difference lies in the fact that annotated lymph node segmentation masks are required to learn a mid-level semantic boundary detector [19], whereas CNN approaches only need LN locations to train [13], [14]. In **abdominal** LN detection, [13] obtains the best trade-off between its CNN model complexity and sampled data configuration. Our best performing CNN model is GoogLeNet-TL (256x256) which obtains an AUC=0.92 and 0.70 TPR/3FP.

Our main difference in dataset preparation protocols from [13] is a more aggressive extraction of random views within a much larger range of FOVs The usage of larger FOVs to capture more image spatial context is inspired by deep zoom-out features [27] that improve semantic segmentation. This image sampling scheme demonstrates contributes to our best reported performance results in both mediastinal LN detection (in this paper) and automated pancreas segmentation [37]. As shown in Figure 1, abdominal LNs are surrounded by many other similar looking objects. Meanwhile, mediastinal LNs are more easily distinguishable due to the images' larger spatial contexts. Finally, from the perspective of the data-model trade-off: "*Do We Need More Training Data or Better Models?*" [29], more abdomen CT scans from distinct patient populations need to be acquired and annotated, in order to take full advantage of deep CNN models of high capacity. Nevertheless, deeper and wider CNN models (e.g., GoogLeNet-RI-L and GoogLeNet-TL-H versus Cifar-10 [13]) have shown improved results in the mediastinal LN detection.

B. Interstitial Lung Disease Detection

The CNN models evaluated in this experiment are *1)* AlexNet-RI (training from scratch on the ILD dataset with random initialization); *2)* AlexNet-TL (with transfer Learning from [3]); *3)* AlexNet-ImNet: pre-trained ImageNet-CNN model [3] with only the last cost function layer retrained from random initialization, according to the six ILD classes (similarly to [8] but without using additional hand-crafted non-deep feature descriptors: GIST, BoVW and so on); *4)* GoogLeNet-RI (random initialization); *5)* GoogLeNet-TL (GoogLeNet with transfer learning from [18]). All ILD images (patches of 64×64 and CT axial slices of 512×512) are re-sampled to a fixed dimension of 256×256 pixels.

We evaluate the ILD classification task with five-fold CV on patient-level split as it is more informative for real clinical performance LOO. The classification accuracies for interstitial lung disease detection are shown in Table III Two sub-tasks on ILD patch and slice classifications are conducted. In general, patch-level ILD classification is less challenging than slice-level classification, as far more data samples can be sampled from the manually annotated ROIs (up to 100 image patches per ROI), available from [20]. From Table III, all five deep models evaluated obtain comparable results within the range of classification accuracy $\in [0.74, 0.76]$. Their averaged model achieves a slightly better accuracy of 0.79.

F1-scores [22], [23], [32] and the confusion matrix (Table V) for patch-level ILD classification using GoogLeNet-TL under five-fold cross-validation (we denote as Patch-CV5) are also computed. F1-scores are reported on patch classification only (32×32 pixel patches extracted from manual ROIs) [22], [23], [32], as shown in Table IV. Both [22], [23] use the evaluation protocol of "leave-one-patient-out" (LOO) protocol, which is arguably much easier and not directly comparable to 10-fold CV [32] or our Patch-CV5. In this study, we classify six ILD classes by adding a consolidation (CD) class to five classes of healthy (normal - NM), emphysema (EM), ground glass (GG), fibrosis (FB), and micronodules (MN) in [22], [23], [32]. Patch-CV10 [32] and Patch-CV5 report similar medium to high F-scores, thus implying that the ILD dataset (although one of the mainstream public medical image datasets) may not adequately represent ILD disease CT lung imaging patterns over a population of only 120 patients. Patch-CV5 Yields higher F-scores than [32] and classifies the extra consolidation (CD) class. The most pressing task at present is to drastically expand the dataset or explore across-dataset deep learning on the union of ILD and LTRC patients [38].

Up to this day, Gao et al. [24] have argued that a new CADe protocol on holistic classification of ILD diseases directly, using axial CT slice attenuation patterns and CNN, may be more realistic for clinical applications. We refer this as slice-level classification, as image patch sampling from manual ROIs can be completely avoided (thus no manual ROI inputs). The experimental results in [24] are conducted with a patient-level hard split of 100 (training) and 20 (testing). The method's testing F-scores (i.e., Slice-Test) are given in Table IV. Note that the F-scores in [24] are not directly comparable to our results due to different evaluation criteria. Only Slice-Test is evaluated and reported in [24] and we find F-scores can change drastically from different rounds of the five-fold CV.

Though a more practical CADe scheme, slice-level CNN learning [24] is very challenging, as it is restricted to a mere 905 CT image slices with tagged ILD labels. We only benchmark the slice-level ILD classification results in this section.

Method	AlexNet-ImNet	AlexNet-RI	AlexNet-TL	GoogLeNet-RI	GoogLeNet-TL	Avg-All
Slice-CV5	0.45	0.44	0.46	0.41	**0.57**	0.53
Patch-CV5	0.76	0.74	0.76	0.75	0.76	**0.79**

TABLE III
COMPARISON OF INTERSTITIAL LUNG DISEASE CLASSIFICATION ACCURACIES ON BOTH SLICE-LEVEL (SLICE-CV5) AND PATCH-BASED (PATCH-CV5) CLASSIFICATION USING FIVE-FOLD CV. BOLD NUMBERS INDICATE THE BEST PERFORMANCE VALUES ON CLASSIFICATION ACCURACY.

Even with the help of data augmentation (described in Sec. II), the classification accuracy of GoogLeNet-TL from Table III is only 0.57. However, transfer learning from ImageNet pre-trained model is consistently beneficial, as evidenced by AlexNet-TL (0.46) versus AlexNet-RI (0.44), and GoogLeNet-TL (0.57) versus GoogLeNet-RI (0.41). It especially prevents GoogLeNet from over-fitting on the limited CADe datasets. Finally, when the cross-validation is conducted by randomly splitting the set of all 905 CT axial slices into five folds, markedly higher F-scores are observed (Slice-Random in Table IV). This further validates the claim that the dataset poorly generalizes ILDs for different patients.

V. ANALYSIS VIA CNN VISUALIZATION

In this section we analyze, via CNN visualization, on the reasons for which transfer learning is beneficial to achieve better performance on CAD applications.

Thoracoabdominal LN Detection. In Figure 10, the first layer convolution filters from five different CNN architectures are visualized. We notice that without transfer learning [35], [5], somewhat blurry filters are learned (AlexNet-RI (256x256), AlexNet-RI (64x64), GoogLeNet-RI (256x256) and GoogLeNet-RI (64x64)). However, in AlexNet-TL (256x256), many higher orders of contrast- or edge-preserving patterns (enable to capture image appearance details) are evidently learned through fine-tuning from ImageNet. With a smaller input resolution, AlexNet-RI (64x64) and GoogLeNet-RI (64x64) can learn image contrast filters to some degree, whereas GoogLeNet-RI (256x256) and AlexNet-RI (256x256) have over-smooth low-level filters throughout.

ILD classification. We focus on analyzing visualization CNN optimization traces and activations with ILD dataset, as its slice-level setting is most similar to ImageNet. Indeed, both use full-size images. The traces of training/validation loss and validation accuracy of AlexNet-RI and AlexNet-TL, are shown in Figure 9. We begin the optimization problem – that of fine-tuning the ImageNet pre-trained CNN to classify a comprehensive set of images – by initializing the parameters near the global optimum. One could deem this process analogous to making adults learn to classify ILDs, as opposed to babies. During the process, the validation loss, having remained at lower values throughout, achieves better final accuracy than the validation loss on a similar problem with random initialization. Meanwhile, the training losses in both cases decrease to near zero indicating that both AlexNet-RI and AlexNet-TL over-fit on ILD dataset, due to its small instance size. The quantitative results in Table III shows that AlexNet-TL and GoogLeNet-TL have consistently better classification accuracies than AlexNet-RI and GoogLeNet-RI, respectively.

The last pooling layer (pool-5) activation maps of the ImageNet pre-trained AlexNet [3] (analogical to AlexNet-ImNet) and AlexNet-TL, obtained by processing two input images of Figure 2 (b,c), are shown in Figure 11 (a,b). The last pooling layer activation map summarizes the entire input image by highlighting where and which relative locations or neural reception fields relative to the image are activated. There are a total of 256 (6x6) reception fields in AlexNet [3]. Pooling units where the relative image location of the disease region is present in the image are highlighted with green boxes. Next, we reconstruct the original ILD images using the process of de-convolution, back-propagating with convolution and un-pooling from the activation maps of the picked pooling units [39]. From the reconstructed images (Figure 11 Bottom). We observe that with fine-tuning, AlexNet-TL detects and localizes objects of interest (ILD disease regions depicted in in Figure 2 (b) and (c)) better than AlexNet-ImNet.

VI. FINDINGS AND FUTURE DIRECTIONS

We summarize our findings as follows.

- Deep CNN architectures in 8, even 22 layers [3], [18] can be useful even for CADe problems where the available training datasets are limited. Previously, CNN models used in medical image analysis applications have often been $2 \sim 5$ orders of magnitude smaller.
- The trade-off between using better learning models and using more training data [29] should be carefully considered for finding an optimal solution of any CADe problem (e.g., mediastinal and abdominal LN detection).
- Limited datasets can be the bottleneck to further advancement of CADe. Building progressively growing (in scales) well annotated datasets is at least as crucial as developing new algorithms. As an analogy in computer vision, the scene recognition problem has made tremendous progress, thanks to the steady and continuous development of Scene-15, MIT Indoor-67, SUN-397 and Place datasets [36].
- Transfer learning from the large scale annotated natural image datasets (ImageNet) to CADe problems has been consistently beneficial in our experiments. This sheds some light on cross-dataset CNN learning in the medical image domain, e.g., the union of ILD [20] and LTRC datasets [38], as suggested in this paper.
- Finally, applications of off-the-shelf deep CNN image features to CADe problems can be improved by either exploring the performance-complementary properties of hand-crafted features [9], [8], [11]; or by training CNNs from scratch and better fine-tuning CNNs on the target medical image dataset, as evaluated in this paper.

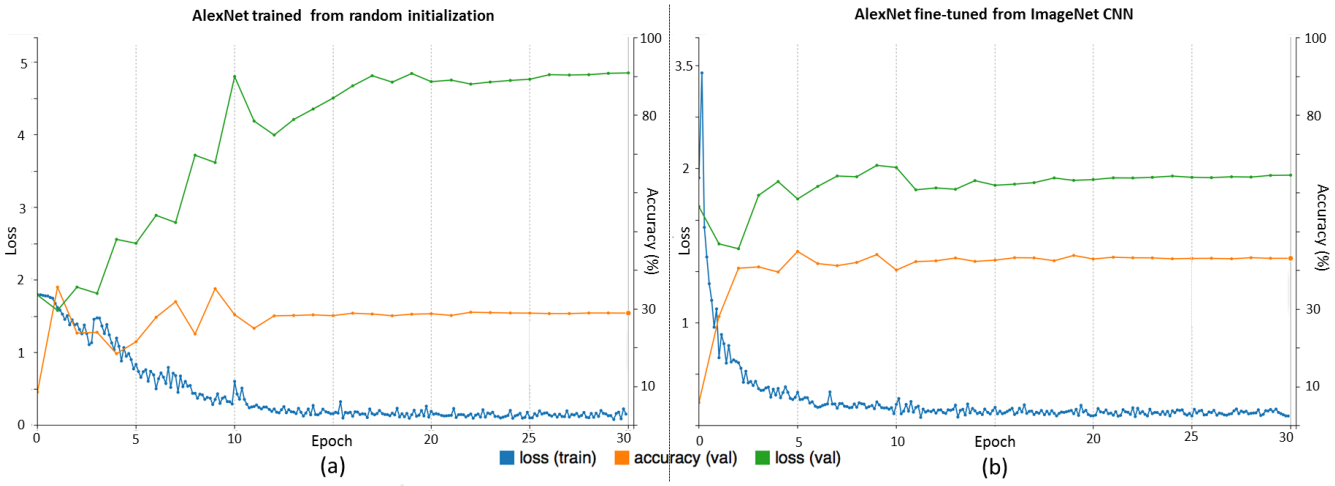


Fig. 9. Traces of training and validation loss (blue and green lines) and validation accuracy (orange lines) during (a) training AlexNet from random initialization and (b) fine-tuning from ImageNet pre-trained CNN, for ILD classification.

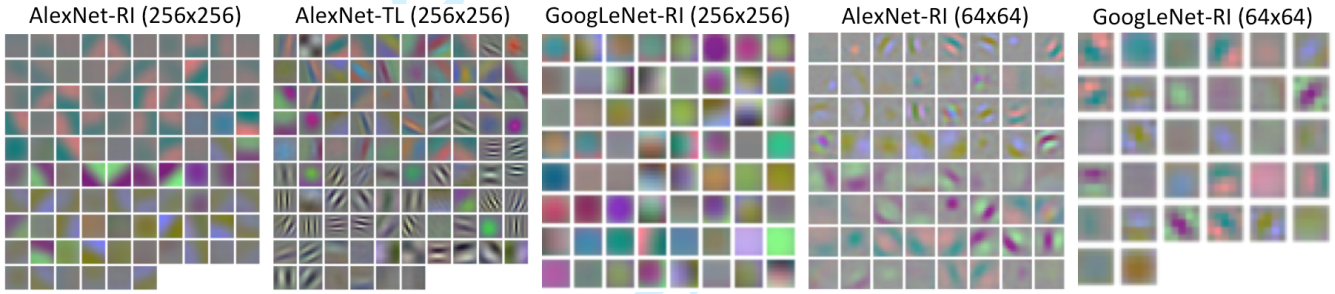


Fig. 10. Visualization of first layer convolution filters of CNNs trained on abdominal and mediastinal LNs in RGB color, from random initialization (AlexNet-RI (256x256), AlexNet-RI (64x64), GoogLeNet-RI (256x256) and GoogLeNet-RI (64x64)) with transfer learning (AlexNet-TL (256x256)).

VII. Conclusion

In this paper, we exploit and extensively evaluate three important, previously under-studied factors on deep convolutional neural networks (CNN) architecture, dataset characteristics, and transfer learning. We evaluate on two different computer-aided diagnosis applications of thoraco-abdominal lymph node detection and interstitial lung disease classification. The empirical evaluation, CNN model visualization, analysis performed in this study, and insights provided can be generalized for designing high performance CAD systems for other medical imaging tasks.

Acknowledgment

This work was supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center, and in part by a grant from the KRIBB Research Initiative Program (Korean Biomedical Scientist Fellowship Program), Korea Research Institute of Bioscience and Biotechnology, Republic of Korea. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). We thank NVIDIA for the K40 GPU donation.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[4] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *Master's Thesis*. University of Toronto, 2009.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and semantic segmentation," *IEEE Trans. on Pat. Anal. and Mach. Intell.*, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. on Pat. Anal. Mach. Intell.*, 2015.

[7] M. Everingham, S. M. A. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.

[8] B. van Ginneken, A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *IEEE ISBI*, 2015, pp. 286–289.

[9] Y. Bar, I. Diamant, H. Greenspan, and L. Wolf, "Chest pathology detection using deep learning with non-medical training," in *ISBI*, 2015.

[10] H. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers, "Interleaved text/image deep mining on a large-scale radiology image database," in *IEEE CVPR*, 2015, pp. 1–10.

[11] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. Scholten, M. Oudkerk, P. de Jong, M. Prokop, and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box," *Medical Image Analysis*, 2015.

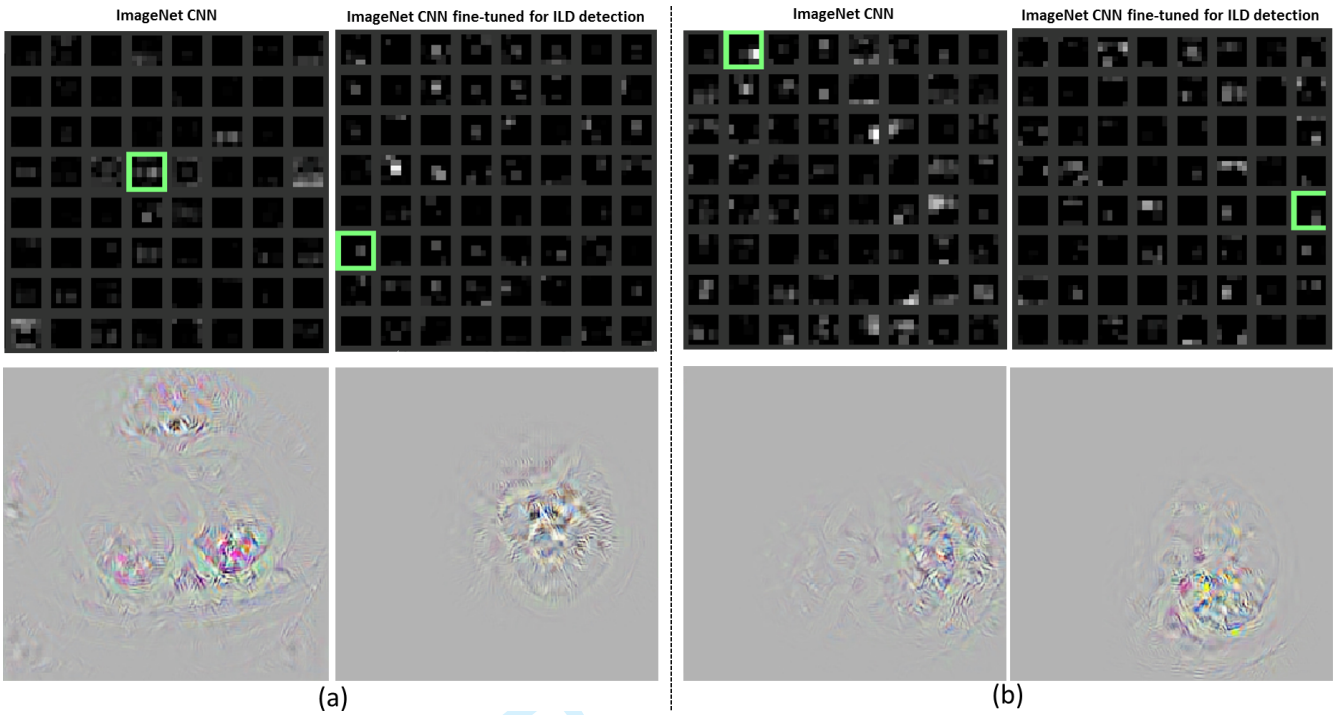


Fig. 11. Visualization of the last pooling layer (pool-5) activations (top). Pooling units where the relative image location of the disease region is located in the image are highlighted with green boxes. The original images reconstructed from the units are shown in the bottom [39]. The examples of (a) and (b) are computed from the input ILD images in Figure 2 (b) and (c), respectively.

[12] O. Russakovsky, J. Deng, and et al., "Imagenet large scale visual recognition challenge," *arXiv:1409.0575*, 2014.
[13] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *MICCAI*, 2014, pp. 520–527.
[14] H. R. Roth, L. J. Lu, Le and, J. Yao, A. Seff, K. M. Cherry, E. Turkbey, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," in *IEEE Trans. on Medical Imaging, to appear*, 2015.
[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, vol. 1, 2005, pp. 886–893.
[17] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *IEEE Conf. on CVPR*. IEEE, 2008.
[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, D. Anguelov, D. Erhan, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
[19] A. Seff, L. Lu, A. Barbu, H. Roth, H.-C. Shin, and R. M. Summers, "Leveraging mid-level semantic boundary cues for computer-aided lymph node detection," in *MICCAI*, 2015.
[20] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *CMIG*, vol. 36, no. 3, pp. 227–238, 2012.
[21] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.
[22] Y. Song, W. Cai, Y. Zhou, and D. D. Feng, "Feature-based image patch approximation for lung tissue classification," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 4, pp. 797–808, 2013.
[23] Y. Song, W. Cai, H. Huang, Y. Zhou, D. Feng, Y. Wang, M. Fulham, and M. Chen, "Large margin local estimate with applications to medical image classification." *IEEE transactions on medical imaging*, 2015.
[24] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, Z. G. Papadakis, A. Depeursinge, M. R. Summers, Z. Xu, and J. D. Mollura, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," in *MICCAI first Workshop on Deep Learning in Medical Image Analysis*, 2015.
[25] 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 *MICCAI*, 2014, pp. 544–552.
[26] L. Lu, M. Liu, X. Ye, S. Yu, and H. Huang, "Coarse-to-fine classification via parametric and nonparametric models for computer-aided diagnosis," in *Proc. ACM Conf. on CIKM*, 2011, pp. 2509–2512.
[27] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," *arXiv preprint arXiv:1412.0774*, 2014.
[28] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, P. Yushkevich *et al.*, "Multi-atlas segmentation with joint label fusion," *Pat. Anal. and Mach. Intel., IEEE Trans. on*, vol. 35, no. 3, pp. 611–623, 2013.
[29] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do we need more training data or better models for object detection?" in *BMVC*, 2012.
[30] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *MICCAI*, 2013.
[31] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.
[32] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *ICARCV*. IEEE, 2014.
[33] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," *http://caffe. berkeleyvision. org*, 2013.
[34] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[35] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPRW*, 2014.
[36] B. Zhou, A. Lapedriza, J. Xiao, and A. Torralba, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.
[37] H. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI*, 2015.
[38] D. Holmes III, B. Bartholmai, R. Karwoski, V. Zavaletta, and R. Robb, "The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease," in *2006 MICCAI Open Science Workshop*, 2006.
[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.