Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2017 May 01.

Published in final edited form as:

IEEE Trans Med Imaging. 2016 May ; 35(5): 1285–1298. doi:10.1109/TMI.2016.2528162.

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

Shin Hoo-Chang [Member, IEEE], **Holger R. Roth [hoochang.shin@nih.gov]**, **Mingchen Gao**, **Le Lu [Senior Member, IEEE]**, **Ziyue Xu**, **Isabella Nogues**, **Jianhua Yao**, **Daniel Mollura**, and **Ronald M. Summers**

Shin Hoo-Chang: hoochang.shin@nih.gov; Le Lu: le.lu.@nih.gov; Ronald M. Summers: rms@nih.gov

Hoo-Chang Shin, Holger R. Roth, Le Lu, Isabella Nogues, 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, Le Lu, Jianhua Yao and Ronald M. Summers are also with Clinical Image Processing Service, Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA

Abstract

Remarkable progress has been made in image recognition, primarily due to the availability of large-scale annotated datasets (i.e. ImageNet) and the revival of deep convolutional neural networks (CNN). CNNs enable learning data-driven, highly representative, layered hierarchical image features from sufficient training data. However, obtaining datasets as comprehensively annotated as ImageNet in the medical imaging domain remains a challenge. There are currently three major techniques that successfully employ CNNs to medical image classification: training the CNN from scratch, using off-the-shelf pre-trained CNN features, and conducting unsupervised CNN pre-training with supervised fine-tuning. Another effective method is transfer learning, i.e., fine-tuning CNN models (supervised) pre-trained from natural image dataset to medical image tasks (although domain transfer between two medical image datasets is also possible).

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 influence of dataset scale and spatial image context on performance. Finally, we examine when and why transfer learning from pre-trained ImageNet (via fine-tuning) can be useful. We study two specific computeraided detection (CADe) problems, namely thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification. We achieve the state-of-the-art performance on the mediastinal LN detection, with 85% sensitivity at 3 false positive per patient, and report the first five-fold cross-validation classification results on predicting axial CT slices with ILD categories. Our

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Correspondence to: Ronald M. Summers, `rms@nih.gov`.

Holger Roth and Mingchen Gao contributed equally to this work.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

extensive empirical evaluation, CNN model analysis and valuable insights can be extended to the design of 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], [2]) and the recent revival of deep convolutional neural networks (CNN) [3], [4]. For data-driven learning, large-scale well-annotated datasets with representative data distribution characteristics are crucial to learning more accurate or generalizable models [5], [4]. Unlike previous image datasets used in computer vision, ImageNet [1] offers 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 [6], [7], e.g., PASCAL [8] and medical image categorization [9], [10], [11], [12]. However, there exists no large-scale annotated medical image dataset comparable to ImageNet, as data acquisition is difficult, and quality annotation is costly.

There are currently three major techniques that successfully employ CNNs to medical image classification: 1) training the "CNN from scratch" [13], [14], [15], [16], [17]; 2) using "off-the-shelf CNN" features (without retraining the CNN) as complementary information channels to existing hand-crafted image features, for Chest X-rays [10] and CT lung nodule identification [9], [12]; and *3*) performing unsupervised pre-training on natural or medical images and fine-tuning on medical target images using CNN or other types of deep learning models [18], [19], [20], [21]. A decompositional 2.5D view resampling and an aggregation of random view classification scores are used to eliminate the "curse-of-dimensionality" issue in [22], in order to acquire a sufficient number of training image samples.

Previous studies have analyzed three-dimensional patch creation for LN detection [23], [24], atlas creation from chest CT [25] and the extraction of multi-level image features [26], [27]. At present, there are several extensions or variations of the decompositional view representation introduced in [22], [28], such as: using a novel vessel-aligned multi-planar image representation for pulmonary embolism detection [29], fusing unregistered multiview for mammogram analysis [16] and classifying pulmonary perifissural nodules via an ensemble of 2D views [12].

Although natural images and medical images differ significantly, conventional image descriptors developed for object recognition in natural images, such as the scale-invariant feature transform (SIFT) [30] and the histogram of oriented gradients (HOG) [31], have been widely used for object detection and 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[10], [9], [12]. They have yielded the best performance results by integrating low-level image features (e.g., GIST [32], bag of visual words (BoVW) and bag-of-frequency [12]). However, the fine-tuning of an ImageNet pre-trained CNN model on medical image datasets has not yet been exploited.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 varying in width (ranging from 5 thousand to 160 million parameters) and depth (various numbers of layers), describe the effects 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 and adapting rich hierarchical image features [5], [33] from the large-scale ImageNet dataset for computer aided diagnosis (CAD). We also explore CNN architectures of the most studied seven-layered "AlexNet-CNN" [4], a shallower "Cifar-CNN" [22], and a much deeper version of "GoogLeNet-CNN" [33] (with our modifications on CNN structures). This study is partially motivated by recent studies [34], [35] in computer vision. The thorough quantitative analysis and evaluation on deep CNN [34] or sparsity image coding methods [35] elucidate the emerging techniques of the time and provide useful suggestions for their future stages of development, respectively.

Two specific computer-aided detection (CADe) problems, namely thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification are studied in this work. On mediastinal LN detection, we surpass all currently reported results. We obtain 86% sensitivity on 3 false positives (FP) per patient, versus the prior state-of-art sensitivities of 78% [36] (stacked shallow learning) and 70% [22] (CNN), as prior state-of-the-art. For the first time, ILD classification results under the patient-level five-fold cross-validation protocol (CV5) are investigated and reported. The ILD dataset [37] contains 905 annotated image slices with 120 patients and 6 ILD labels. Such sparsely annotated datasets are generally difficult for CNN learning, due to the paucity of labeled instances.

Evaluation protocols and details are critical to deriving significant empirical findings [34]. Our experimental results suggest that different CNN architectures and dataset resampling protocols are critical for the LN detection tasks where the amount of labeled training data is sufficient and spatial contexts are local. Since LN images are more flexible than ILD images with respect to resampling and reformatting, LN datasets may be more readily augmented by such image transformations. As a result, LN datasets contain more training and testing data instances (due to data auugmentation) than ILD datasets. They nonetheless remain less comprehensive than natural image datasets, such as ImageNet. Fine-tuning ImageNet-trained models for ILD classification is clearly advantageous and yields early promising results, when the amount of labeled training data is highly insufficient and multiclass 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 CNNs that merely use off-the-shelf CNN features, 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

We employ CNNs (with the characteristics defined above) to thoraco-abdominal lymph node (LN) detection (evaluated separately on the mediastinal and abdominal regions) and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

interstitial lung disease (ILD) detection. For LN detection, we use randomly sampled 2.5D views in CT [22]. We use 2D CT slices [38], [39], [40] for ILD detection. We then evaluate and compare CNN performance results.

Until the detection aggregation approach [22], [41], thoracoabdominal lymph node (LN) detection via CADe mechanisms has yielded poor performance results. In [22], each 3D LN candidate 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 [22], using a "Cifar-10" CNN. Using the thoracoabdominal LN detection datasets [22], *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) comprises more than 150 lung diseases affecting the interstitium, which can severely impair the patient's ability to breathe. Gao et al. [40] 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: *a*/ sampling patches within the 2D ROIs (Regions of Interest provided by [37]), then *b*/ classifying patches into seven category labels (six disease labels and one "''healthy"" label). Song et al. [38], [39] only address the second sub-task of patch-level classification under the "leave-one-patient-out" (LOO) criterion. By training on the moderate-to-small scale ILD dataset [37], *our main objective is to exploit and benchmark CNN based ILD classification performances under the CV5 metric (which is more realistic and unbiased than LOO [38], [39] and hard-split [40]), with and without transfer learning*.

Thoracoabdominal Lymph Node Datasets

We use the publicly available dataset from [22], [41]. 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 [22], where positive and negative LN candidates are sampled with the fields-of-view (FOVs) of 30mm to 45mm, surrounding the annotated and detected LN centers (obtained by a candidate generation process). More precisely, [22], [41], [36] follow a coarse-to-fine CADe scheme, partially inspired by [42], which operates with ~ 100% detection recalls at the cost of approximately 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 close to the total number of positives. LN candidates are randomly extracted from fields-of-view (FOVs) spanning 35mm to 128mm in soft-tissue window [−100, 200HU]. This allows us to capture multiple spatial scales of image context [43], [44]). The samples are then rescaled to a 64 × 64 pixel resolution via B-spline interpolation. A few examples of LNs with axial, coronal, and sagittal views encoded in RGB color images [22] are shown in Figure 1.

Unlike the heart or the liver, lymph nodes have no predetermined anatomic orientation. Hence, the purely random image resampling (with respect to scale, displacement and orientation) and reformatting (the axial, coronal, and sagittal views are in any system

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

randomly resampled coordinates) is a natural choice, which also happens to yield high CNN performance. Although we integrate three channels of information from three orthogonal views for LN detection, the pixel-wise spatial correlations between or among channels are not necessary. The convolutional kernels in the lower level CNN architectures can learn the optimal weights to linearly combine the observations from the axial, coronal, and sagittal channels by computing their dot-products. Transforming axial, coronal, and sagittal representations to RGB also facilitates transfer learning from CNN models trained on ImageNet.

This learning representation (i.e., "built-in CNN") is flexible, in that it naturally combines multiple sources or channels of information. In the recent literature [45], even heterogeneous class-conditional probability maps can be combined with raw images to improve performance. This set-up is similar to that of other works in computer vision, such as [46], where heterogeneous image information channels are jointly fed into the CNN convolutional layers for high-accuracy human parsing and segmentation. Finally, if there are correlations among CNN input channels, one may observe the corresponding correlated patterns in the learned filters.

In summary, the assumption that there are or must be pixel-wise spatial correlations among input channels does not apply to the CNN model representation. For other medical imaging problems, such as pulmonary embolism detection [29], in which orientation can be constrained along the attached vessel axis, vessel-aligned multi-planar image representation (MPR) is more effective than randomly aligned MPR.

Interstitial Lung Disease Dataset

We utilize the publicly available dataset of [37]. It contains 905 image slices from 120 patients, with six lung tissue types annotations containing at least one of the following: healthy (NM), emphysema (EM), ground glass (GG), fibrosis (FB), micronodules (MN) and consolidation (CD) (Figure 3). At the slice level, the objective is to classify the status of "presence/absence" of any of the six ILD classes for an input axial CT slice [40]. Characterizing an arbitrary CT slice against any possible ILD type, without any manual ROI (in contrast to [38], [39]), 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 by randomly sampling several 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.

In the publicly available ILD dataset, very few CT slices are labeled as normal or healthy. The remaining CT slices cannot be simply classified as normal, because many ILD disease regions or slices have not yet been labeled. ILD [37] is a partially labeled database; this is one of its main limitations. Research is being conducted to address this issue. In particular,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[47] has proposed to fully label the ILD dataset pixel-wise via proposed segmentation label propagation.

To leverage the CNN architectures designed for color images and to transfer CNN parameters pre-trained on ImageNet, we transform all gray-scale axial CT slice images via three CT window ranges: lung window range [−1400, −200HU], high-attenuation range [−160, 240HU], and low-attenuation range[−1400; −950HU]. We then encode the transformed images into RGB channels (to be aligned with the input channels of CNN models [4], [33] pre-trained from natural image datasets [1]). The low-attenuation CT window is useful for visualizing certain texture patterns of lung diseases (especially emphysema). The usage of different CT attenuation channels improves classification results over the usage of a single CT windowing channel, as demonstrated in [40]. More importantly, these CT windowing processes do not depend on the lung segmentation, which instead is directly defined in the CT HU space. Figure 4 shows a representative example of lung, high-attenuation, and low-attenuation CT windowing for an axis lung CT slice.

As observed in [40], lung segmentation is crucial to holistic slice-level ILD classification. We empirically compare performance in two scenarios with a rough lung segmentation[1] There is no significant difference between two setups. Due to the high precision of CNN based image processing, highly accurate lung segmentation is not necessary. The localization of ILD regions within the lung is simultaneously learned through selectively weighted CNN reception fields in the deepest convolutional layers during the classification based CNN training [49], [50]. Some areas outside of the lung appear in both healthy or diseased images. CNN training learns to ignore them by setting very small filter weights around the corresponding regions (Figure 13). This observation is validated by [40].

III. Methods

In this study, we explore, evaluate and analyze the influence of various CNN Architectures, dataset characteristics (when we need more training data or better models for object detection [51]) 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 mainly explore three convolutional neural network architectures (CifarNet [5], [22], AlexNet [4] and GoogLeNet [33]) with different model training parameter values. The current deep learning models [22], [52], [53] in medical image tasks are at least 2 ~ 5 orders of magnitude smaller than even AlexNet [4]. More complex CNN models [22], [52] have only about 150K or 15K parameters. Roth et al. [22] adopt the CNN architecture tailored to the Cifar-10 dataset [5] and operate on image windows of 32 × 32 × 3 pixels for lymph node

[1]This can be achieved by segmenting the lung using simple label-fusion methods [48]. In the first case, we overlay the target image slice with the average lung mask among the training folds. In the second, we perform simple morphology operations to obtain the lung boundary. In order to retain information from the inside of the lung, we apply Gaussian smoothing to the regions outside of the lung boundary.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

detection, while the simplest CNN in [54] has only one convolutional, pooling, and FC layer, respectively.

We use CifarNet [5] as used in [22] as a baseline for the LN detection. AlexNet [4] and GoogLeNet [33] are also modified to evaluate these state-of-the-art CNN architecture from ImageNet classification task [2] to our CADe problems and datasets. A simplified illustration of three CNN architectures exploited is shown in Figure 5. CifarNet always takes 32 × 32 × 3 image patches as input while AlexNet and GoogLeNet are originally designed for the fixed image dimension of 256 × 256 × 3 pixels. We also reduced the filter size, stride and pooling parameters of AlexNet and GoogLeNet to accommodate a smaller input size of 64 × 64 × 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 (32 × 32) or CifarNet (dropping 32 × 32); AlexNet (256 × 256) or AlexNet-H (high resolution); AlexNet (64 × 64) or AlexNet-L (low resolution); GoogLeNet (256 × 256) or GoogLeNet-H and GoogLeNet (64 × 64) or GoogLeNet-L (dropping 3 since all image inputs are three channels).

a) CifarNet—CifarNet, introduced in [5], was the state-of-the-art model for object recognition on the Cifar10 dataset, which consists of 32 × 32 images of 10 object classes. The objects are normally centered in the images. Some example images and class categories from the Cifar10 dataset are shown in Figure 7. CifarNet has three convolution layers, three pooling layers, and one fully-connected layer. This CNN architecture, also used in [22] 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 [4], achieved significantly improved performance over the other non-deep learning methods for ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. This success has revived the interest in CNNs [3] 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 approximately 60 million free parameters. AlexNet is our default CNN architecture for evaluation and analysis in the remainder of the paper.

c) GoogLeNet—The GoogLeNet model proposed in [33], 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 dimensions into a single new filter (refer to Figure 6). 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 6. 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. ImageNet: Large Scale Annotated Natural Image Dataset

ImageNet [1] has more than 1.2 million 256 × 256 images categorized under 1000 object class categories. There are more than 1000 training images per class. The database is organized according to the WordNet [55] hierarchy, which currently contains only nouns in 1000 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. ImageNet 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 classes, the objects belonging to each ImageNet class category can be occluded, partial and small, relative to those in the previous public image datasets. This significant intra-class variation poses greater challenges to any data-driven learning system that builds a classifier to fit given data and generalize to unseen data. For comparison, some example images of Cifar10 dataset and ImageNet images in the “tennis ball” class category are shown in Figure 7. 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. Training Protocols and Transfer Learning

When **learned from scratch**, all the parameters of CNN models are initialized with 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 [56] and NVidia K40 GPUs to train the CNNs.

AlexNet and GoogLeNet CNN models can be either learned from scratch or **fine-tuned from pre-trained models**. Gir-shick et al. [6] find that, by applying ImageNet pre-trained ALexNet to PASCAL dataset [8], 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, [57], [58] 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For transfer learning, we follow the approach of [57], [6] where all CNN 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 random initialization or transfer learning as AlexNet-RI and AlexNet-TL, and GoogLeNet-RI and GoogLeNet-TL. We found that the transfer learning strategy yields the best performance results. Determining the optimal learning rate for different layers is challenging, especially for very deep networks such as GoogLeNet.

We also perform experiments using "**off-the-shelf**" CNN features of AlexNet pre-trained on ImageNet and training only the final classifier layer to complete the new CADe classification tasks. Parameters in the convolutional and fully connected layers are fixed and are used as deep image extractors, as in [10], [9], [12]. We refer to this model as AlexNet-ImNet in the remainder of the paper. Note that [10], [9], [12] train support vector machines and random forest classifiers using ImageNet pre-trained CNN features. Our simplified implementation is intended to determine whether fine-tuning the "end-to-end" CNN network is necessary to improve performance, as opposed to merely training the final classification layer. This is a slight modification from the method described in [10], [9], [12].

Finally, transfer learning in CNN representation, as empirically verified in previous literature [59], [60], [61], [11], [62], can be effective in various cross-modality imaging settings (RGB images to depth images [59], [60], natural images to general CT and MRI images [11], and natural images to neuroimaging [61] or ultrasound [62] data). More thorough theoretical studies on cross-modality imaging statistics and transferability will be needed for future studies.

IV. Evaluations and Discussions

In this section, we evaluate and compare the performances of nine CNN model configurations (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 [22], [41], [37].

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 by averaging probabilities, as in [22].

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-L, 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 reduce the stride from 4 to 2. For the modified GoogLeNet-RI (64 × 64), we decrease the number of first layer convolution filters from 64

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to 32, the pad size from 3 to 2, the kernel size from 7 to 5, stride from 2 to 1 and the stride of the subsequent pooling layer from 2 to 1. We slightly reduce the number of convolutional filters in order to accommodate the smaller input image sizes of target medical image datasets [22], [37], while preventing over-fitting. This eventually improves performance on patch-based classification. CifarNet is used in [22] to detect LN samples of 32 × 32 × 3 images. For consistency purposes, we down-sample 64 × 64 × 3 resolution LN sample images to the dimension of 32 × 32 × 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. Of the nine investigated CNN models, CifarNet, AlexNet-ImNet and GoogLeNet-RI-H generally yielded the least competitive detection accuracy results. Our LN datasets are significantly more complex (i.e., display much larger within-class appearance variations), especially due to the extracted fields-of-view (FOVs) of (35mm–128mm) compared to (30mm–45mm) in [22], 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 adequate for mediastinal and abdominal LN detection tasks. To complement “off-the-shelf” CNN features, [10], [9], [12] 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 with 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). This indicates that transfer learning offers a much better initialization of CNN parameters than random initialization. Likewise, AlexNet-TL-H consistently outperforms AlexNet-RI-H, though by 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 11.

GoogLeNet-TL-H yields results similar to AlexNet-TL-H’s for the mediastinal LN detection, and slightly outperforms Alex-Net-H for abdominal LN detection. AlexNet-RI-H exhibits 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 [36] for **mediastinal** LN detection: GoogLeNet-RI-L obtains an AUC=0.95 and 0.85 TPR/3FP, versus AUC=0.92 and 0.70 TPR/3FP [22] and 0.78

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TPR/3FP [36] 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 [36], whereas CNN approaches only need LN locations for training [22]. In **abdominal** LN detection, [22] obtains the best tradeoff between its CNN model complexity and sampled data configuration. Our best performing CNN model is GoogLeNet-TL (256 × 256) which obtains an AUC=0.92 and 0.70 TPR/3FP.

The main difference between our dataset preparation protocol and that from [22] 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 [44] that improve semantic segmentation. This image sampling scheme contributes to our best reported performance results in both mediastinal LN detection (in this paper) and automated pancreas segmentation [45]. 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*?" [51], 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 [22]) have shown improved results in the mediastinal LN detection.

Figure 9 provides examples of misclassified lymph nodes (in axial view) (both false negatives (**Left**) and false positives(**Right**)), from the Abdomen and Mediastinum datasets. The overall reported LN detection results are clinically significant, as indicated in [63].

B. Interstitial Lung Disease Classification

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 [4]); *3*) AlexNet-ImNet: pre-trained ImageNet-CNN model [4] with only the last cost function layer retrained from random initialization, according to the six ILD classes (similar to [9] but without using additional hand-crafted non-deep feature descriptors, such as GIST and BoVW); *4*) GoogLeNet-RI (random initialization); *5*) GoogLeNet-TL (GoogLeNet with transfer learning from [33]). 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 than LOO. The classification accuracy rates 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 [37]. From Table III, all five deep models evaluated obtain comparable results within the range of classification accuracy rates [0.74, 0.76]. Their averaged model achieves a slightly better accuracy of 0.79.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

F1-scores [38], [39], [54] 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) [38], [39], [54], as shown in Table IV. Both [38] and [39] use the evaluation protocol of "leave-one-patient-out" (LOO), which is arguably much easier and not directly comparable to 10-fold CV [54] 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 [38], [39], [54]. Patch-CV10 [54] and Patch-CV5 report similar medium to high F-scores. This implies 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 [54] and classifies the extra consolidation (CD) class. At present, the most pressing task is to drastically expand the dataset or to explore across-dataset deep learning on the combined ILD and LTRC datasets [64].

Recently, Gao et al. [40] 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 to this as slice-level classification, as image patch sampling from manual ROIs can be completely avoided (hence, no manual ROI inputs will be provided). The experimental results in [40] 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 [40] are not directly comparable to our results, due to different evaluation criteria. Only Slice-Test is evaluated and reported in [40], and we find that F-scores can change drastically from different rounds of the five-fold CV.

While it is a more practical CADe scheme, slice-level CNN learning [40] is very challenging, as it is restricted to only 905 CT image slices with tagged ILD labels. We only benchmark the slice-level ILD classification results in this section. 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 obtained (Slice-Random in Table IV). This further validates the claim that the dataset poorly generalizes ILDs for different patients. Figure 10 shows examples of misclassified ILD patches (in axial view), with their ground truth labels and inaccurately classified labels.

No existing work has reached the performance requirements for a realistic clinical setting [40], in which simple ROI-guided image patch extraction and classification (which requires manual ROI selection by clinicians) is implemented. The main goal of this paper is to investigate the three factors (CNN architectures, dataset characteristics and transfer learning) that affect performance on a specific medical image analysis problem and to ultimately deliver clinically relevant results. For ILD classification, the most critical performance bottlenecks are the challenge of cross-dataset learning and the limited patient population

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

size. We attempt to overcome these obstacles by merging the ILD [37] and LTRC datasets. Although the ILD [37] and LTRC datasets [64] (used in [19]) were generated and annotated separately, they contain many common disease labels. For instance, the ILD disease classes emphysema (EM), ground glass (GG), fibrosis (FB), and micronodules (MN) belong to both datasets, and thus can be jointly trained/tested to form a larger and unified dataset.

Adapting fully convolutional CNN or FCNN to parse every pixel location in the ILD lung CT images or slices, or adapting other methods from CNN based semantic image segmentation using PASCAL or ImageNet, may improve accuracy and efficiency. However, current FCNN approaches [65], [66] lack adequate spatial resolution in their directly output label space. A segmentation label propagation method was recently proposed [47] to provide full pixel-wise labeling of the ILD data images. In this work, we sample image patches from the slice using the ROIs for the ILD provided in the dataset, in order to be consistent with previous methods in patch-level [38], [39], [54] and slice-level classification [40].

C. Evaluation of Five CNN Models using ILD Classification

In this work, we mainly focus on AlexNet and GoogLeNet. AlexNet is the first notably successful CNN architecture on the ImageNet challenge and has rekindled significant research interests on CNN. GoogLeNet is the state-of-the-art deep model, which has outperformed other notable models, such as AlexNet, OverFeat, and VGGNet [67], [68] in various computer vision benchmarks. Likewise, a reasonable assumption is that OverFeat and VGGNet may generate quantitative performance results ranked between AlexNet's and GoogLeNet's. For completeness, we include the Overfeat and VGGNet in the following evaluations, to bolster our hypothesis.

d) Overfeat—OverFeat is described in [67] as an integrated framework for using CNN for classification, localization and detection. Its architecture is similar to that of AlexNet, but contains far more parameters (e.g., 1024 convolution filters in both "conv4" and "conv5" layers compared to 384 and 256 convolution kernels in the "conv4" and "conv5" layers of AlexNet), and operates more densely (e.g., smaller kernel size of 2 in "pool2" layer "pool5" compared to the kernel size 3 in "pool2" and "pool5" of AlexNet) on the input image. Overfeat is the winning model of the ILSVRC 2013 in detection and classification tasks.

e) VGGNet—The VGGNet architecture is introduced in [68], where it is designed to significantly increase the depth of the existing CNN architectures with 16 or 19 layers. Very small 3 × 3 size convolutional filters are used in all convolution layers with a convolutional stride of size 1, in order to reduce the number of parameters in deeper networks. Since VGGNet is substantially deeper than the other CNN models, VGGNet is more susceptible to the vanishing gradient problem [69], [70], [71]. Hence, the network may be more difficult to train. Training the network requires far more memory and computation time than AlexNet. We use the 16 layer variant as our default VGGNet model in our study.

The classification accuracy results for ILD slice and patch level classification of five CNN architectures (CifarNet, AlexNet, Overfeat, VGGNet and GoogLeNet) are shown in Table VI. Based on the analysis in Sec. IV-B, transfer learning is only used for the slice level classification task. From Table VI, quantitative classification accuracy rates increase as the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CNN model becomes more complex (CifarNet, AlexNet, Overfeat, VGGNet and GoogLeNet, in ascending order), for both ILD slice and patch level classification problems. The reported results validate our assumption that OverFeat's and VGGNets performance levels fall between AlexNet's and GoogLeNets (this observation is consistent with the computer vision findings). CifarNet is designed for images with smaller dimensions (32 × 32 images), and thus is not catered to classification tasks involving 256 × 256 images.

To investigate the performance difference between five-fold cross-validation (CV) in Sec. IV-B and leave-one-patient-out (LOO) validation, this experiment is performed under the LOO protocol. By comparing results in Table III (CV-5) to those in Table VI (LOO), one can see that LOOs quantitative performances are remarkably better than CV-5's. For example, in ILD slice-level classification, the accuracy level drastically increases from 0.46 to 0.867 using AlexNet-TL, and from 0.57 to 0.902 for GoogLeNet-TL.

CNN training is implemented with the Caffe [56] deep learning framework, using a NVidia K40 GPU on Ubuntu 14.04 Linux OS. All models are trained for up to 90 epochs with early stopping criteria, where a model snapshot with low validation loss is taken for the final model. Other hyper-parameters are fixed as follows: momentum: 0.9; weight decay: 0.0005; and a step learning rate schedule with base learning rate of 0.01, decreased by a factor of 10 every 30 epochs. The image batch size is set to 128, except for GoogLeNet's (64) and VGG-16's (32), which are the maximum batch sizes that can fit in the NVidia K40 GPU with 12GB of memory capacity. Table VII illustrates the training time and memory requirements of the five CNN architectures on ILD patch-based classification up to 90 epochs.

D. Training with "Equal Prior" vs. "Biased Prior"

Medical datasets are often "biased", in that the number of healthy samples is much larger than the number of diseased instances, or that the numbers of images per class are uneven. In ILD dataset, the number of fibrosis samples is about 3.5 times greater than the number of emphysema samples. The number of non-LNs is 3 ~ 4 times greater than the number of LNs in lymph node detection. Different sampling or resampling rates are routinely applied to both ILD and LN detection to balance the data sample number or scale per class, as in[22]. We refer this as "Equal Prior". If we use the same sampling rate, that will lead to a "Biased Prior" across different classes.

Without loss of generality, after GoogLeNet is trained on the training sets under "Equal" or "Biased" priors, we compare its classification results on the balanced validation sets. Evaluating a classifier on a biased validation set will cause unfair assessment of its performance. For instance, a classifier that predicts every image patch as "non-LN" will still achieve a 70% accuracy rate on a biased set with 3.5 times as many non-LN samples as LN samples. The classification accuracy results of GoogLeNet trained under two configurations are shown in Table VIII. Overall, it achieves lower accuracy results when trained with a "biased prior" in both tasks, and the accuracy difference for ILD patch-based classification is small.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

V. Analysis via CNN Learning Traces & lulVisualization

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

Thoracoabdominal LN Detection

In Figure 12, the first layer convolution filters from five different CNN architectures are visualized. We notice that without transfer learning [57], [6], somewhat blurry filters are learned (AlexNet-RI (256 × 256), AlexNet-RI (64 × 64), GoogLeNet-RI (256 × 256) and GoogLeNet-RI (64 × 64)). However, in AlexNet-TL (256 × 256), many higher orders of contrast- or edge-preserving patterns (that enable capturing image appearance details) are evidently learned through fine-tuning from ImageNet. With a smaller input resolution, AlexNet-RI (64 × 64) and GoogLeNet-RI (64 × 64) can learn image contrast filters to some degree; whereas, GoogLeNet-RI (256 × 256) and AlexNet-RI (256 × 256) have over-smooth low-level filters throughout.

ILD classification

We focus on analyzing visual CNN optimization traces and activations from the ILD dataset, as its slice-level setting is most similar to ImageNet's. Indeed, both datasets use full-size images. The traces of the training loss, validation loss and validation accuracy of AlexNet-RI and AlexNet-TL, are shown in Figure 11. For AlexNet-RI in Figure 11 (a), the training loss significantly decreases as the number of training epochs increases, while the validation loss notably increases and the validation accuracy does not improve much before reaching a plateau. With transfer learning and fine-tuning, much better and consistent performances of training loss, validation loss and validation accuracy traces are obtained (see Figure 11 (b)). 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 close to an optimal solution. One could compare this process to making adults learn to classify ILDs, as opposed to babies. During the process, the validation loss, having remained at lower values throughout, achieves higher final accuracy levels than the validation loss on a similar problem with random initialization. Meanwhile, the training losses in both cases decrease to values near zero. This indicates that both AlexNet-RI and AlexNet-TL over-fit on the ILD dataset, due to its small instance size. The quantitative results in Table III indicate 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 Ima-geNet pre-trained AlexNet [4] (analogical to AlexNet-ImNet) and AlexNet-TL, obtained by processing two input images of Figure 2 (b,c), are shown in Figure 13 (a,b). The last pooling layer activation map summarizes the entire input image by highlighting which relative locations or neural reception fields relative to the image are activated. There are a total of 256 (6 × 6) reception fields in AlexNet [4]. 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 chosen pooling units [72]. From the reconstructed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

images (Figure 13 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. The filters shown in Figure 13 that better localize regions on the input images (Figure 2 (b) and (c)) respectively, produce relatively higher activations (in the top 5%) among all 512 reception field responses in the fine-tuned AlexNet-TL model. As observed in [73], the final CNN classification score can not be driven solely by a single strong activation in the receptions fields, but often by a sparse set of high activations (i.e., varying selective or sparse activations per input image).

VI. Findings and Future Directions

We summarize our findings as follows.

- Deep CNN architectures with 8, even 22 layers [4], [33], 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 ~ 5 orders of magnitude smaller.
- The trade-off between using better learning models and using more training data [51] should be carefully considered when searching for an optimal solution to any CADe problem (e.g., mediastinal and abdominal LN detection).
- Limited datasets can be a bottleneck to further ad-vancement of CADe. Building progressively growing (in scale), well annotated datasets is at least as crucial as developing new algorithms. This has been accomplished, for instance, in the field of computer vision. The well-known 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 [58].
- 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 the ILD [37] and LTRC datasets [64], 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 [10], [9], [12], or by training CNNs from scratch and better fine-tuning CNNs on the target medical image dataset, as evaluated in this paper.

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 CNN performance on two different computer-aided diagnosis applications: thoraco-abdominal lymph node detection and interstitial lung disease classification. The empirical evaluation, CNN model visualization, CNN performance analysis, and conclusive insights can be generalized to the design of high performance CAD systems for other medical imaging tasks.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Acknowledgments

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. Deng J, Dong W, Socher R, Li L-J, Li k, Fei-Fei L. Imagenet: a large-scale hierarchical image database. IEEE CVPR. 2009
2. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L. Imagenet large scale visual recognition challenge. 2014 arXiv: 1409.0575.
3. LeCun Y, Bottou L, Bengio y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE. 1998; 86(11):2278–2324.
4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. NIPS. 2012
5. Krizhevsky, A. Master's Thesis. Dept. of Comp. Science, University of Toronto; 2009. Learning multiple layers of features from tiny images.
6. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2015
7. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015
8. Everingham M, Eslami SMA, Van Gool L, Williams C, Winn J, Zisserman A. The pascal visual object classes challenge: a retrospective. International journal of computer vision. 2015; 111(1):98–136.
9. van Ginneken B, Setio A, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. IEEE ISBI. 2015:286–289.
10. Bar, y; Diamant, I.; Greenspan, H.; Wolf, L. Chest pathology detection using deep learning with non-medical training. IEEE ISBI. 2015
11. Shin, H.; Lu, L.; Kim, L.; Seff, A.; Yao, J.; Summers, R. Interleaved text/image deep mining on a large-scale radiology image database; IEEE Conf. on CVPR; 2015. p. 1-10.
12. Ciompi F, de Hoop B, van Riel SJ, Chung K, Scholten E, Oudkerk M, de Jong P, Prokop M, van Ginneken B. 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
13. Menze B, Reyes M, Van Leemput K. The multimodal brain tumor image segmentation benchmark (brats). Medical Imaging, IEEE Trans. on. 2015; 34(10):1993–2024.
14. Pan Y, Huang W, Lin Z, Zhu W, Zhou J, Wong J, Ding Z. Brain tumor grading based on neural networks and convolutional neural networks. IEEE EMBC. 2015:699–702.
15. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. IPMI. 2015:588–599.
16. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. MIC-CAI. 2015:652–660.
17. Wolterink JM, Leiner T, Viergever MA, Isˇgum I. Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. MICCAI. 2015:589–596.
18. Schlegl, T.; Ofner, J.; Langs, G. Medical Computer Vision: Algorithms for Big Data. Springer; 2014. Unsupervised pre-training across image domains improves lung tissue classification; p. 82-93.
19. Hofmanninger, J.; Langs, G. Mapping visual features to semantic profiles for retrieval in medical imaging; IEEE Conf. on CVPR; 2015.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

20. Carneiro G, Nascimento J. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Trans. Pattern Anal. Mach. Intell. 2013; 35(11):2592–2607. [PubMed: 24051722]
21. Li R, Zhang W, Suk H, Wang L, Li J, Shen D, Ji S. Deep learning based imaging data completion for improved brain disease diagnosis. MICCAI. 2014
22. Roth H, Lu L, Liu J, Yao J, Seff A, Cherry KM, Turkbey E, Summers RM. Improving computer-aided detection using convolu-tional neural networks and random view aggregation. IEEE Trans. on Medical Imaging. 2016
23. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D. Automatic detection and segmentation of lymph nodes from ct data. Medical Imaging, IEEE Trans. on. 2012; 31(2):240–250.
24. Feulner J, Zhou SK, Hammon M, Hornegger J, Comaniciu D. Lymph node detection and segmentation in chest ct data using discriminative learning and a spatial prior. Medical image analysis. 2013; 17(2):254–270. [PubMed: 23246185]
25. Feuerstein M, Glocker B, Kitasaka T, Nakamura Y, Iwano S, Mori K. Mediastinal atlas creation from 3-d chest computed tomography images: application to automated detection and station mapping of lymph nodes. Medical image analysis. 2012; 16(1):63–74. [PubMed: 21641269]
26. Lu, L.; Devarakota, P.; Vikal, S.; Wu, D.; Zheng, Y.; Wolf, M. Medical Computer Vision. Large Data in Medical Imaging. Springer; 2014. Computer aided diagnosis using multilevel image features on large-scale evaluation; p. 161-174.
27. Lu L, Bi J, Wolf M, Salganicoff M. Effective 3d object detection and regression using probabilistic segmentation features in ct images. IEEE CVPR. 2011
28. Lu L, Barbu A, Wolf M, Liang J, Salganicoff M, Comaniciu D. Accurate polyp segmentation for 3d ct colonography using multi-staged probabilistic binary learning and compositional model. IEEE CVPR. 2008
29. Tajbakhsh N, Gotway MB, Liang J. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. MICCAI. 2015
30. Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. 2004; 60(2):91–110.
31. Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE CVPR. 2005; 1:886–893.
32. Torralba A, Fergus R, Weiss Y. Small codes and large image databases for recognition. IEEE CVPR. 2008:1–8.
33. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Rabinovich, A. Going deeper with convolutions; IEEE Conf. on CVPR; 2015.
34. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. BMVC. 2014
35. Chatfield K, Lempitsky VS, Vedaldi A, Zisserman A. The devil is in the details: an evaluation of recent feature encoding methods. BMVC. 2011
36. Seff A, Lu L, Barbu A, Roth H, Shin H-C, Summers RM. Leveraging mid-level semantic boundary cues for computer-aided lymph node detection. MICCAI. 2015
37. Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti P-A, Mu¨ller H. Building a reference multimedia database for interstitial lung diseases. Computerized medical imaging and graphics. 2012; 36(3):227–238. [PubMed: 21803548]
38. Song Y, Cai W, Zhou Y, Feng DD. Feature-based image patch approximation for lung tissue classification. Medical Imaging, IEEE Trans. on. 2013; 32(4):797–808.
39. Song Y, Cai W, Huang H, Zhou Y, Feng D, Wang Y, Fulham M, Chen M. Large margin local estimate with applications to medical image classification. IEEE Trans. on Medical Imaging. 2015
40. Gao M, Bagci U, Lu L, Wu A, Buty M, Shin H-C, Roth H, Papadakis ZG, Depeursinge A, Summers MR, Xu Z, Mollura JD. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. MICCAI first Workshop on Deep Learning in Medical Image Analysis. 2015

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

41. Seff A, Lu L, Cherry KM, Roth HR, Liu J, Wang S, Hoffman J, Turkbey EB, Summers RM. 2d view aggregation for lymph node detection using a shallow hierarchy of linear classifiers. MICCAI. 2014:544–552. [PubMed: 25333161]
42. Lu, L.; Liu, M.; Ye, X.; Yu, S.; Huang, H. Coarse-to-fine classification via parametric and nonparametric models for computer-aided diagnosis. ACM Conf. on CIKM; 2011. p. 2509-2512.
43. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 2013; 35(8):1915–1929. [PubMed: 23787344]
44. Mostajabi M, Yadollahpour P, Shakhnarovich G. Feedforward semantic segmentation with zoom-out features. arXiv preprint. 2014 arXiv:1412.0774.
45. Roth H, Lu L, Farag A, Shin H-C, Liu J, Turkbey E, Summers RM. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. MICCAI. 2015
46. Liang X, Xu C, Shen X, Yang J, Liu S, Tang J, Lin L, Yan S. Human parsing with contextualized convolutional neural network. IEEE ICCV. 2015:1386–1394.
47. Gao M, Lu L, Nogues I, Summers MR, Mollura D. Segmentation label propagation using deep convolutional neural networks and dense conditional random field. IEEE ISBI. IEEE. 2016
48. Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich P, et al. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 2013; 35(3):611–623. [PubMed: 22732662]
49. Oquab M, Bottou L, Laptev I, Sivic J. Is object localization for free?–weakly-supervised learning with convolutional neural networks. IEEE CVPR. 2015:685–694.
50. Oquab M, Bottou L, Laptev I, Josef S. Learning and transferring mid-level image representations using convolutional neural networks. IEEE CVPR. 2015:1717–1724.
51. Zhu X, Vondrick C, Ramanan D, Fowlkes C. Do we need more training data or better models for object detection? BMVC. 2012
52. Ciresan D, Giusti A, Gambardella L, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. MICCAI. 2013
53. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, Shen D. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage. 2015; 108:214–224. [PubMed: 25562829]
54. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. IEEE ICARCV. 2014:844–848.
55. Miller GA. Wordnet: a lexical database for english. Communications of the ACM. 1995; 38(11): 39–41.
56. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick RB, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding.". ACM Multimedia. 2014; 2:4.
57. Razavian AS, Azizpour H, Sullivan J, Carlsson S. Cnn features off-the-shelf: an astounding baseline for recognition. IEEE CVPRW. 2014:512–519.
58. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. NIPS. 2014:487–495.
59. Gupta S, Girshick R, Arbelez P, Malik J. Learning rich features from rgb-d images for object detection and segmentation. ECCV. 2014:345–360.
60. Gupta S, Arbelez P, Girshick R, Malik J. Indoor scene understanding with rgb-d images: Bottom-up segmentation object detection and semantic segmentation. International Journal of Computer Vision. 2015; 112(2):133–149.
61. Gupta A, Ayhan M, Maida A. Natural image bases to represent neuroimaging data. ICML. 2013:987–994.
62. Chen H, Dou Q, Ni D, Cheng J, Qin J, Li S, Heng P. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. MICCAI. 2015:507–514.
63. Kim, L.; Roth, H.; Lu, L.; Wang, S.; Turkbey, E.; Ronald, SM. Performance assessment of retroperitoneal lymph node computer-assisted detection using random forest and deep convolutional neural network learning algorithms in tandem; The 102nd Annual Meeting of Radiological Society of North America; 2014.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

64. Holmes III D, Bartholmai B, Karwoski R, Zavaletta V, Robb R. The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease. MICCAI Open Science Workshop. 2006
65. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE CVPR. 2015
66. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected crfs. ICLR. 2015
67. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, Le-Cun Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR. 2014
68. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR. 2014
69. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998; 6(02):107–116.
70. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural computation. 2006; 18(7):1527–1554. [PubMed: 16764513]
71. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. Neural Networks, IEEE Transactions on. 1994; 5(2):157–166.
72. Zeiler MD, Fergus R. Visualizing and understanding convolu-tional networks. ECCV. 2014:818–833.
73. Agrawal P, Girshick R, Malik J. Analyzing the performance of multilayer neural networks for object recognition. ECCV. 2014

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

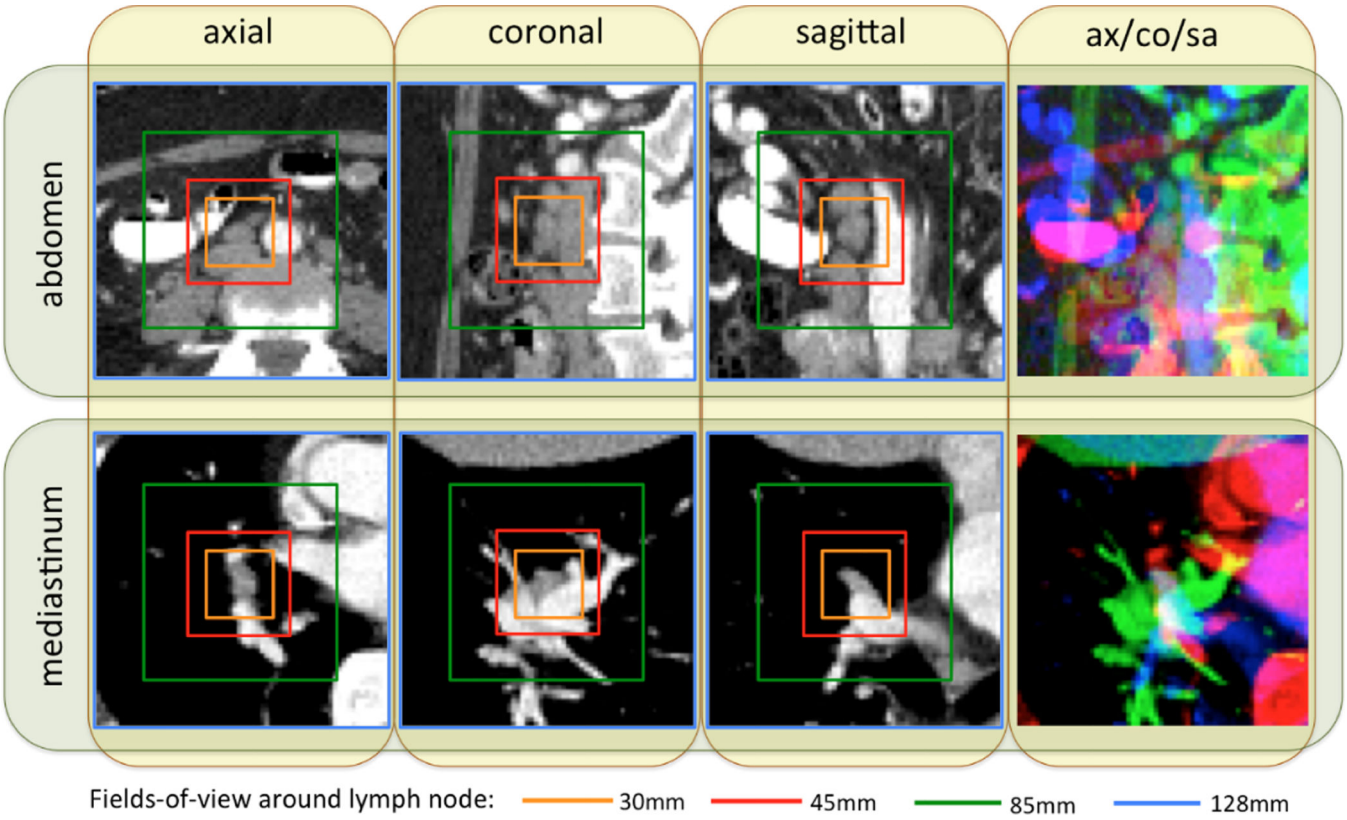

Fig. 1.

Some examples of abdominal and mediastinal lymph nodes sampled on axial (ax), coronal (co), and sagittal (sa) views, with four different fields-of-views (30mm: orange; 45mm: red; 85mm: green; 128mm: blue) surrounding lymph nodes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

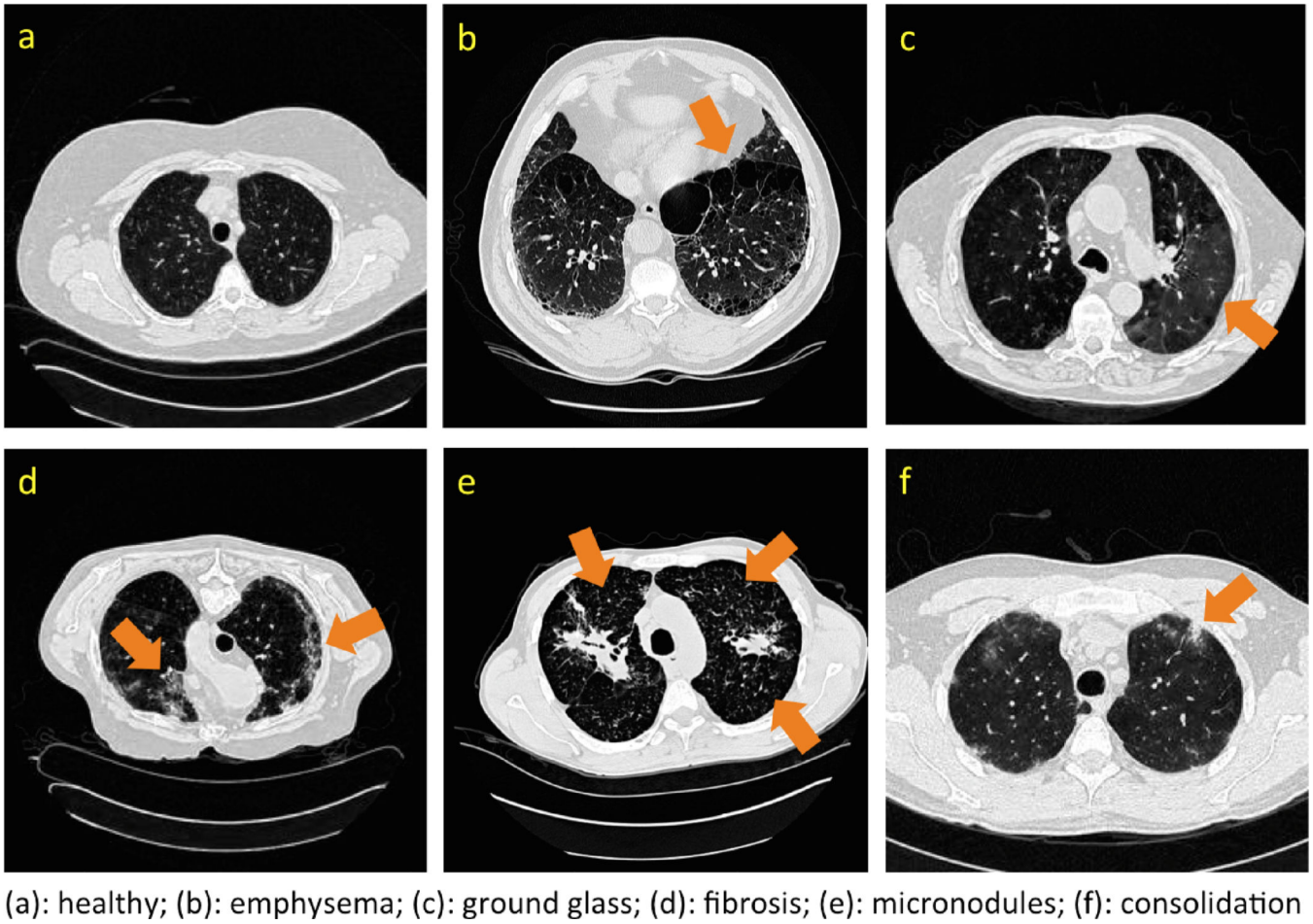

Fig. 2.

Some examples of CT image slices with six lung tissue types in the ILD dataset [37]. Disease tissue types are located with dark orange arrows.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

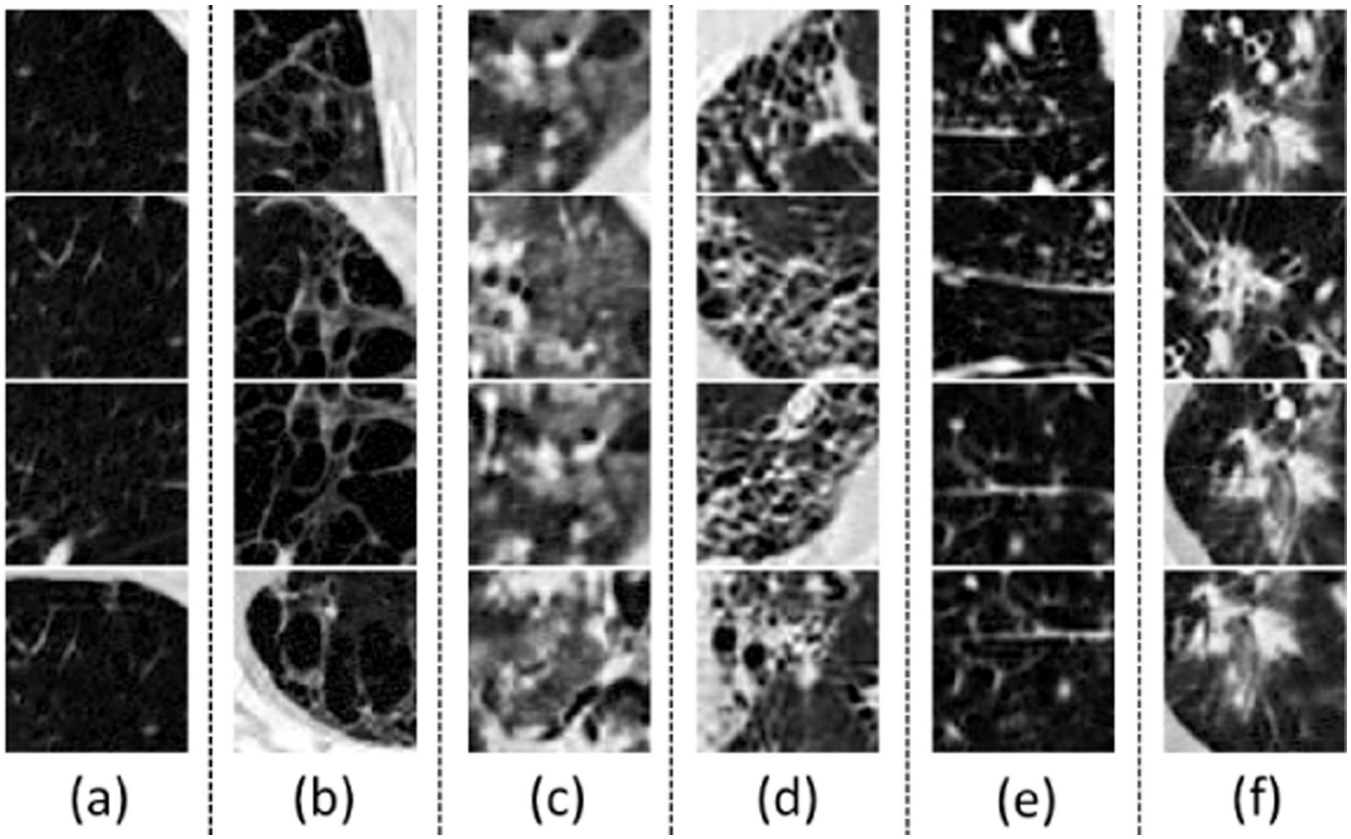

Fig. 3.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

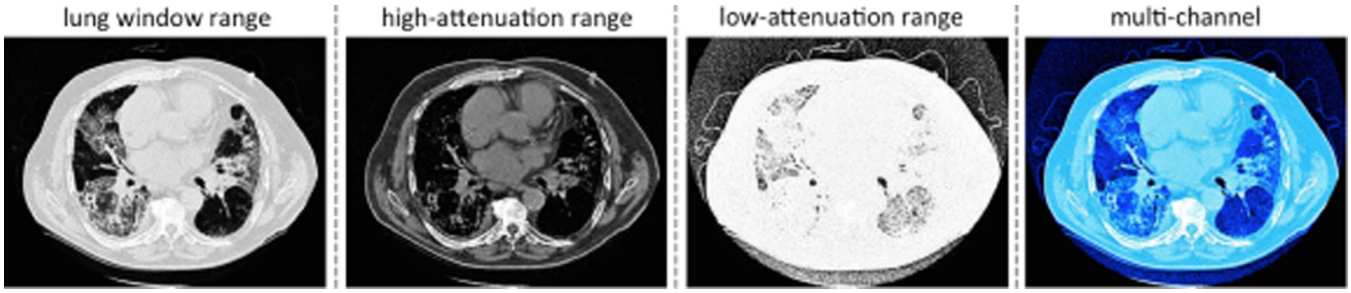

Fig. 4.

An example of lung/high-attenuation/low-attenuation CT windowing for an axis lung CT slice. We encode the lung/high-attenuation/low-attenuation CT windowing into red/green/blue channels.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

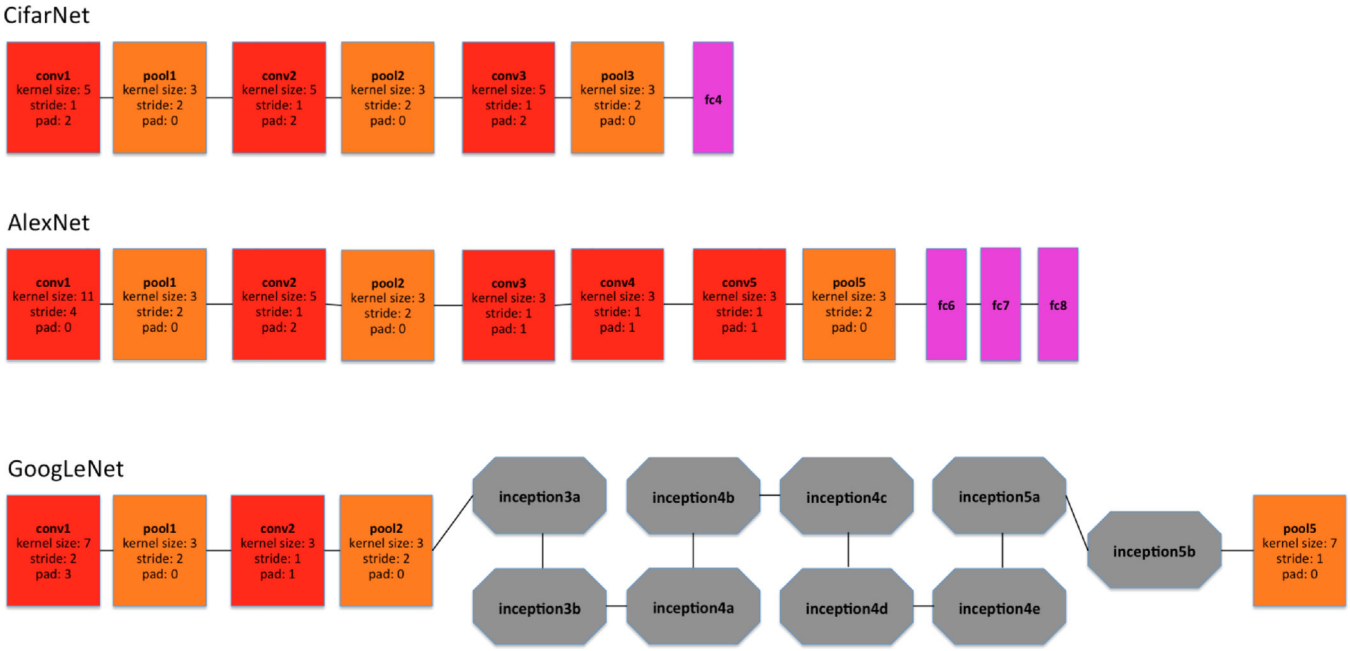

Fig. 5.

A simplified illustration of the CNN architectures used. GoogLeNet [33] contains two convolution layers, three pooling layers, and nine inception layers. Each of the inception layer of GoogLeNet consists of six convolution layers and one pooling layer.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

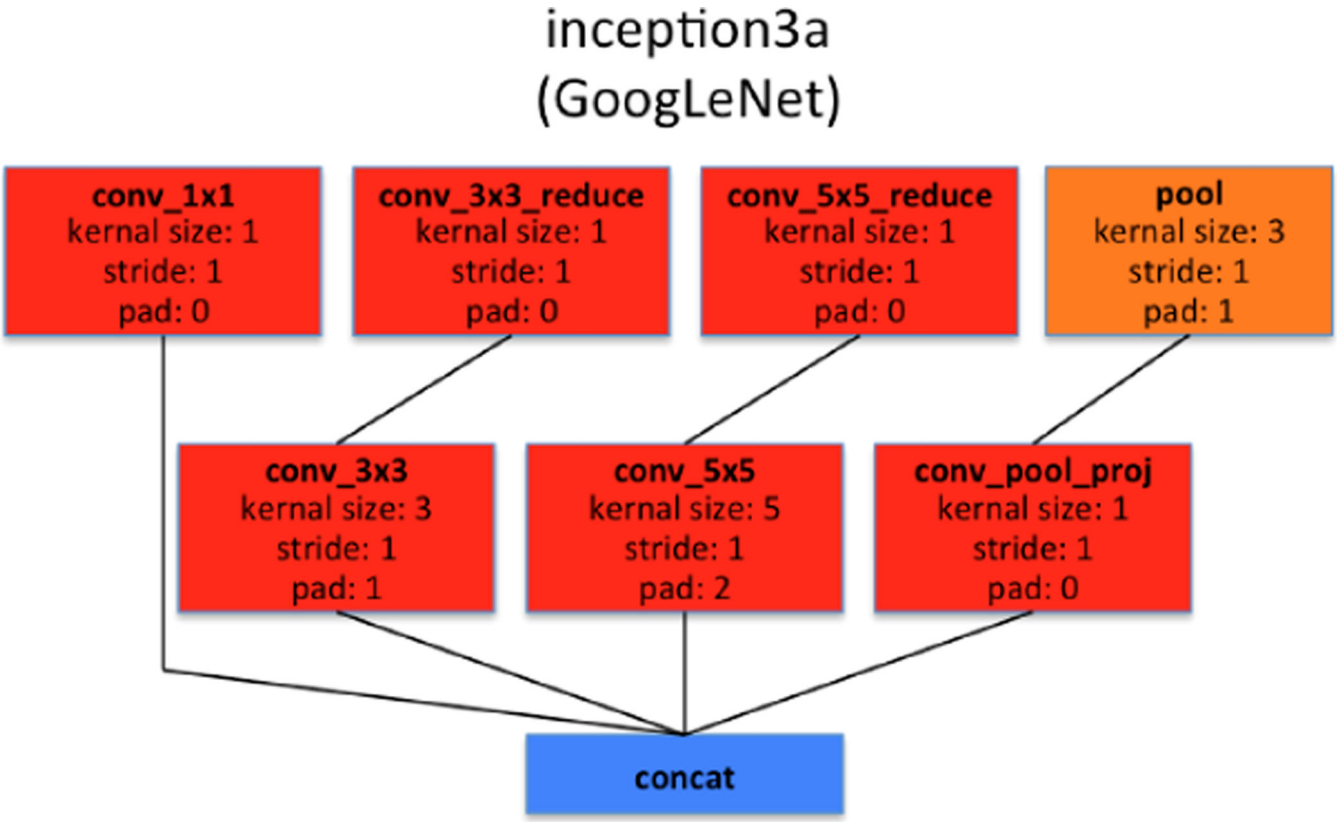

Fig. 6.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

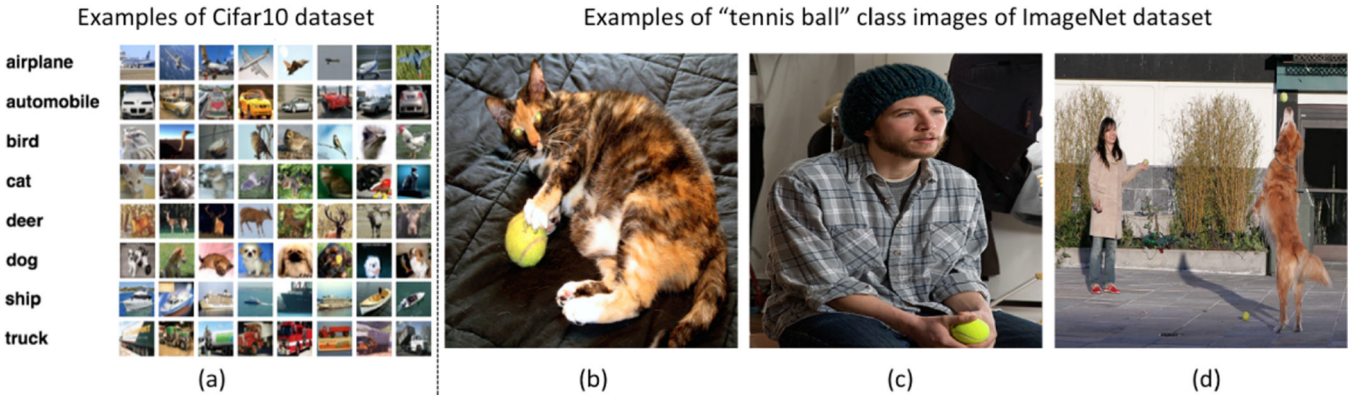

Fig. 7.

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

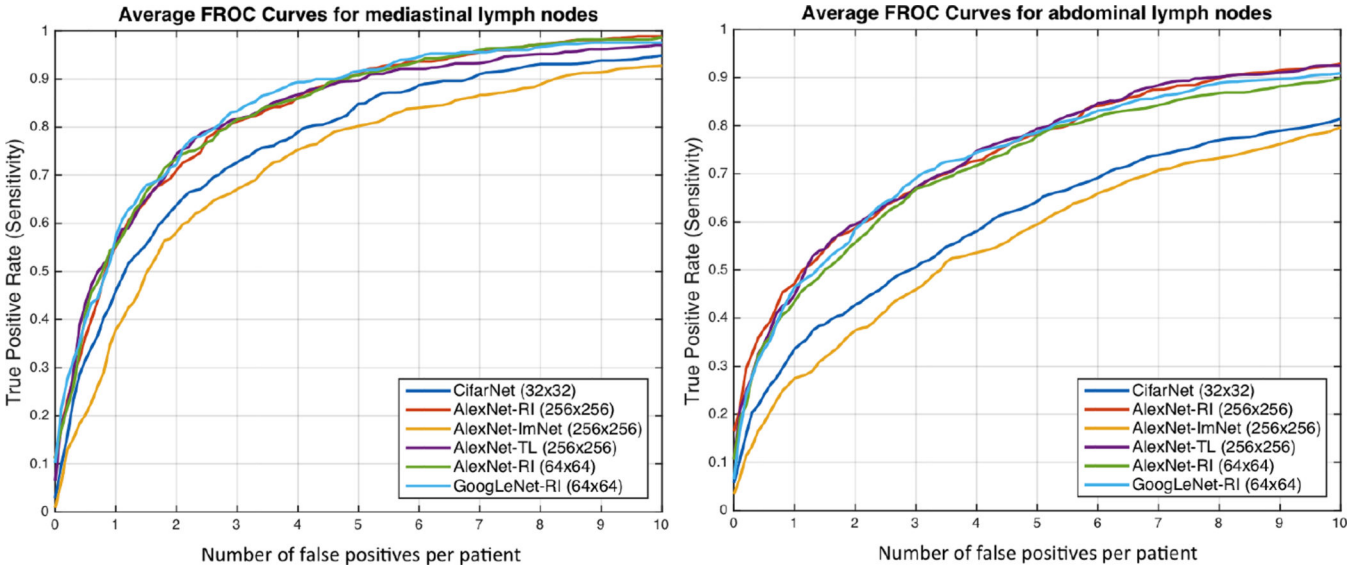

Fig. 8.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

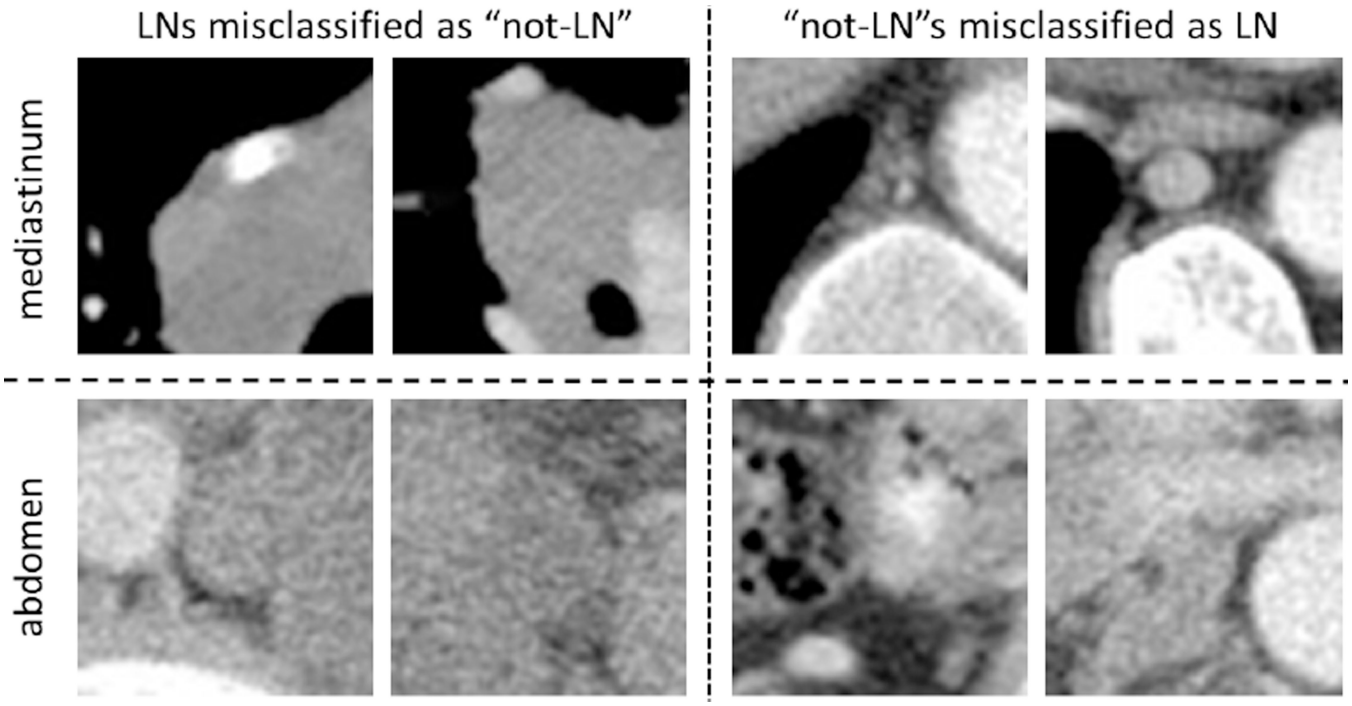

Fig. 9.

Examples of misclassified lymph nodes (in axial view) of both false negatives (**Left**) and false positives (**Right**). Mediastinal LN examples are shown in the **upper** row, and abdominal LN examples in the **bottom** row.

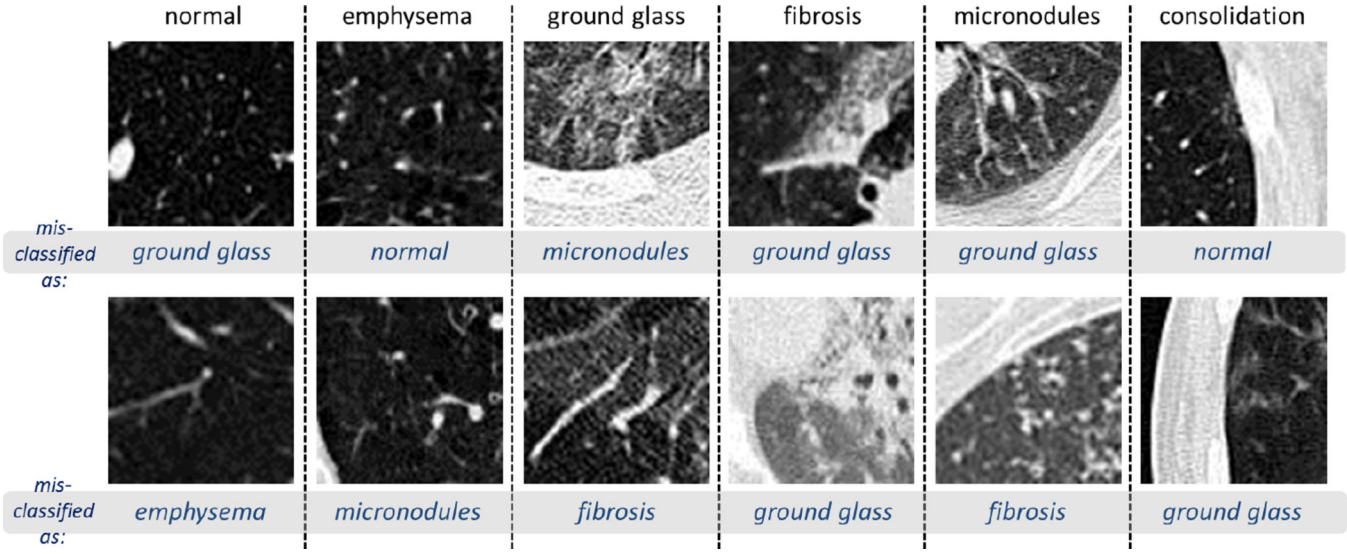

Fig. 10.

Visual examples of misclassified ILD 64 × 64 patches (in axial view), with their ground truth labels and inaccurately classified labels.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

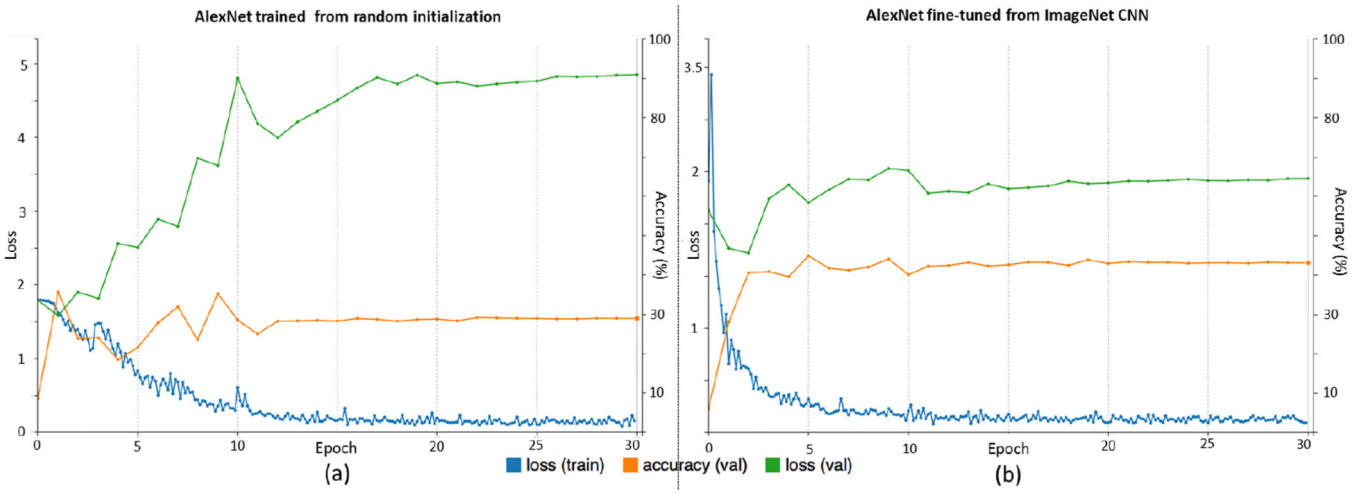

Fig. 11.

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

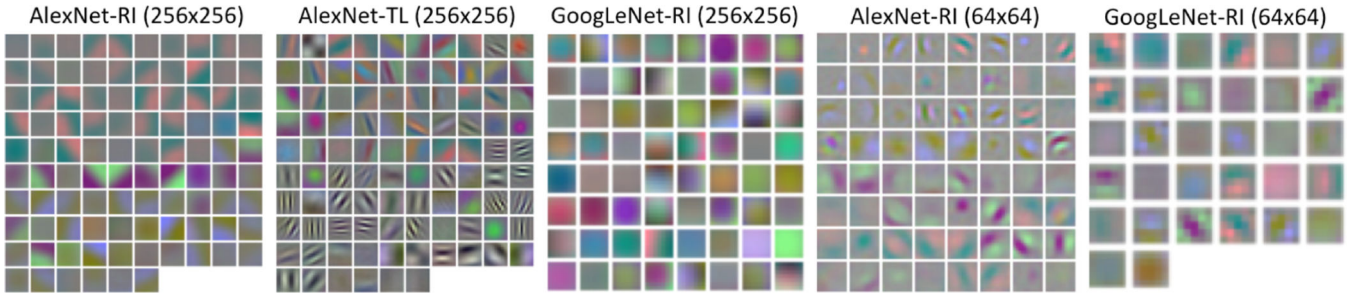

Fig. 12.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

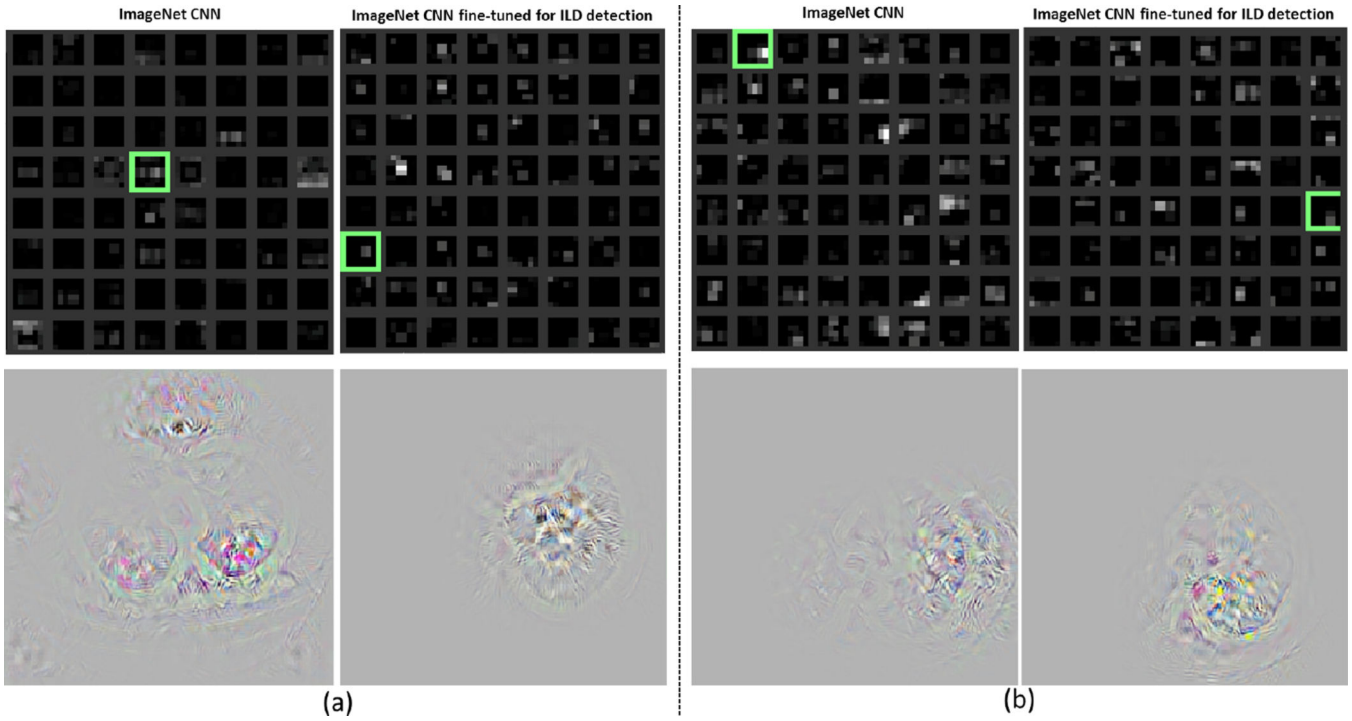

Fig. 13.

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 [72]. The examples in (a) and (b) are computed from the input ILD images in Figure 2 (b) and (c), respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table I

Average Number of Images in Each Fold for Disease Classes, When Dividing the Dataset in 5-Fold Patient Sets.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table II

Comparison of mediastinal and abdominal LN detection results using various CNN models. Numbers in bold indicate the best performance values on classification accuracy.

Region	Mediastinum		Abdomen	
Method	**AUC**	**TPR/3FP**	**AUC**	**TPR/3FP**
[41]	-	0.63	-	0.70
[22]	**0.92**	0.70	**0.94**	**0.83**
[36]	-	**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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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.

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**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table IV

Comparison of interstitial lung disease classification results using F-Scores: NM, EM, GG, FB, MN and CD.

	NM	EM	GG	FB	MN	CD
Patch-LOO [38]	0.84	0.75	0.78	0.84	0.86	-
Patch-LOO [39]	0.88	0.77	0.80	0.87	0.89	-
Patch-CV10 [54]	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 [40]	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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table V

Confusion matrix for ILD classification (Patch-level) with five-fold CV using GoogLeNet-TL.

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**

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table VI

Classification results on ILD and LN detection with LOO.

Method	ILD-Slice	Method	ILD-Patch
CifarNet	-	CifarNet	0.799
AlexNet-TL	0.867	AlexNet-TL	0.865
Overfeat-TL	0.877	Overfeat-TL	0.879
VGG-16-TL	0.90	VGG-16-TL	0.893
GoogLeNet-TL	0.902	GoogLeNet-TL	0.911

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table VII

Training time and Memory Requirements of the five CNN architectures on ILD Patch-Based Classification up to 90 epochs.

	CifarNet	AlexNet	Overfeat	VGG-16	GoogLeNet
Time	7m16s	1h2m	1h26m	20h24m	2h49m
Memory	2.25 GB	3.45 GB	4.22 GB	9.26 GB	5.37 GB

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table VIII

Classification accuracies for ILD Slice and LN patch-level detection with “equal prior“ AND “biased prior“, using GoogLeNet-TL.

	ILD-Slice	ILD-Patch
Equal Prior	0.902	0.953
Biased Prior	0.872	0.952