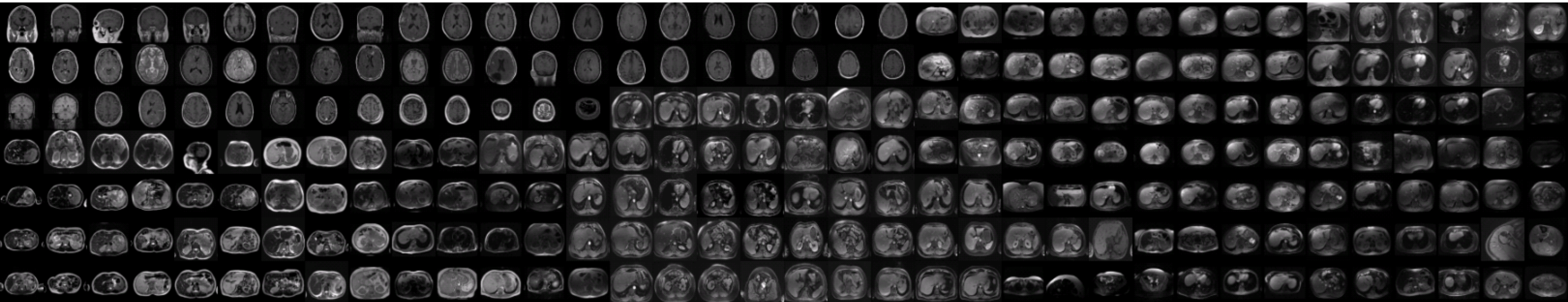
National Institutes of Health
Turning Discovery Into Health

Building Truly Large-Scale Medical Image Databases: Deep Label Discovery and Open-Ended Recognition (GTC 2017, S7595)

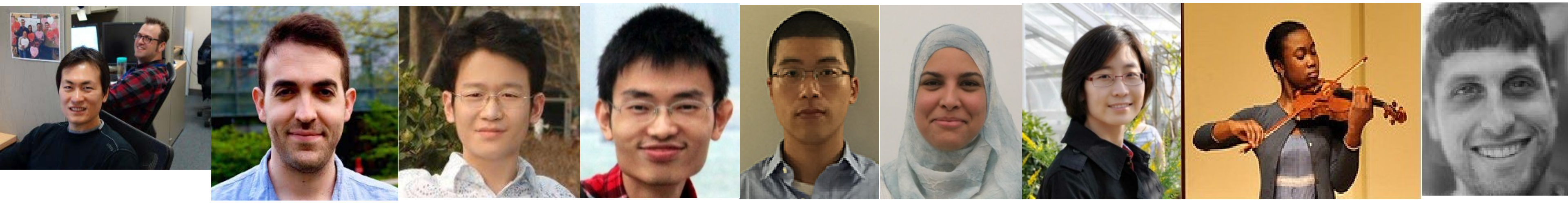
Le Lu, PhD, Staff Scientist, le.lu@nih.gov;

NIH Clinical Center, Radiology and Imaging Sciences

5/11/2017



NIH National Institutes of Health
Turning Discovery Into Health



Q1: Do deep learning and deep neural networks help in medical imaging or medical image analysis problems? (Yes)

- ✓ **Deep CAD:** Lymph node application package (52.9% → 85%, 83%) and many CAD Applications
- ✓ **Deep Segmentation → Precision Medicine in Radiology & Oncology:** Pancreas segmentation application package (~53% → 81.14% in Dice Coefficient) and beyond (prostate segmentation, …)
- ✓ **Deep Lung** (Interstitial Lung Disease) Application **Package** + **DL Reading Chest X-ray**; **Pathological Lung Segmentation**, …
- ✓ **Unsupervised category discovery using looped deep pseudo-task optimization (mapping large-scale radiology database with category meta-labels) → Learning from PACS!**
- ✓ **A large-scale Chest X-ray database (with NLP based annotation): Dataset and Benchmark**

- **Updates & Publications can be downloaded: www.cs.jhu.edu/~lelu; https://clinicalcenter.nih.gov/drd/staff/le_lu.html**

Perspectives

- Why the previous or current computer-aided diagnosis (CADx) systems are not particularly successful yet? **Integrating machine decisions is not easy for human doctors**: Good doctors hate to use; bad doctors are confused and do not know how to use? --> **Human-machine collaborative decision making process**
  - Make machine decision more **interpretable** is very critical for the collaborative system --> learning mid-level attributes or embedding?

- **Preventive** medicine: what human doctors cannot do (in very large scales: millions of general population, at least not economical): → **first-reader population risk profiling** …?

- **Precision** Medicine: a) **new imaging biomarkers** in precision medicine to **better assist human doctors to make more precise decisions;** b) **patient-level similarity retrieval system** for personalized diagnosis/therapy treatment: show by examples!

Three Key Problems (I)

Computer-aided Detection (CADe) and Diagnosis (CADx)

- Lung, Colon pre-cancer detection; Bone and Vessel imaging (6 years of industrial R&D at Siemens Corporation and Healthcare, 10+ product transfer; 13 conference papers in CVPR/ECCV/ICCV/MICCAI/WACV/CIKM, 12 US/EU patents, 27 Inventions)
- **Lymph node**, colon polyp, bone lesion detection using Deep CNN + Random View Aggregation (TMI 2016a; MICCAI 2014a)
- Empirical analysis on **Lymph node detection** and **interstitial lung disease** (ILD) classification using CNN (TMI 2016b)
- Non-deep models for CADe using **compositional** representation (MICCAI 2014b) and **+mid-level cues** (MICCAI 2015b); **deep regression** based **multi-label** ILD prediction (*in submission*); missing label issue in ILD (ISBI 2016); ISBI 2017 …

➢ **Clinical Impacts**: producing various ***high performance*** "second or first reader" **CAD use cases** and applications → effective imaging based prescreening (triage) tools on a cloud based platform for large population

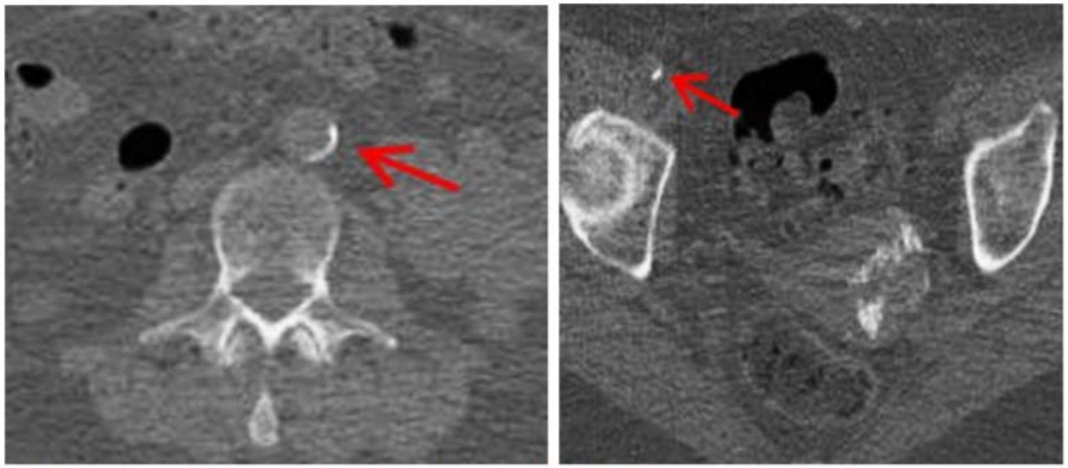
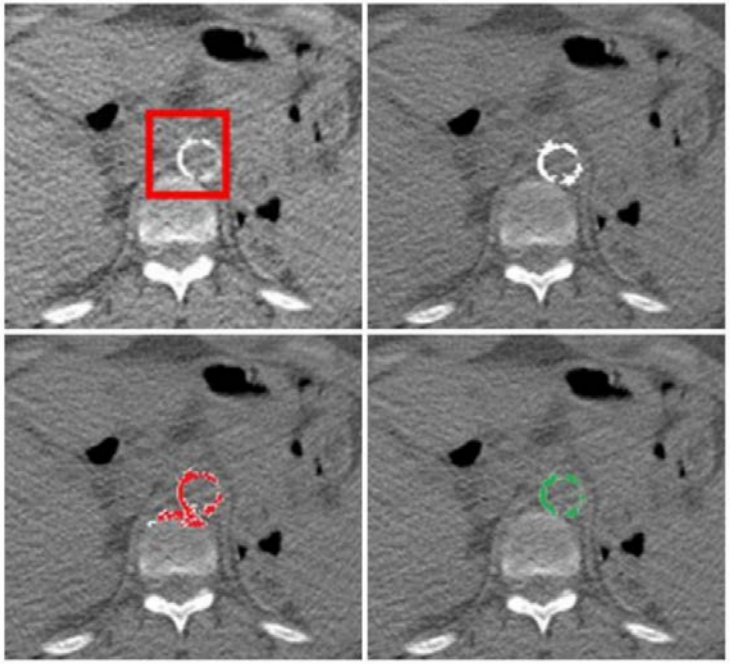
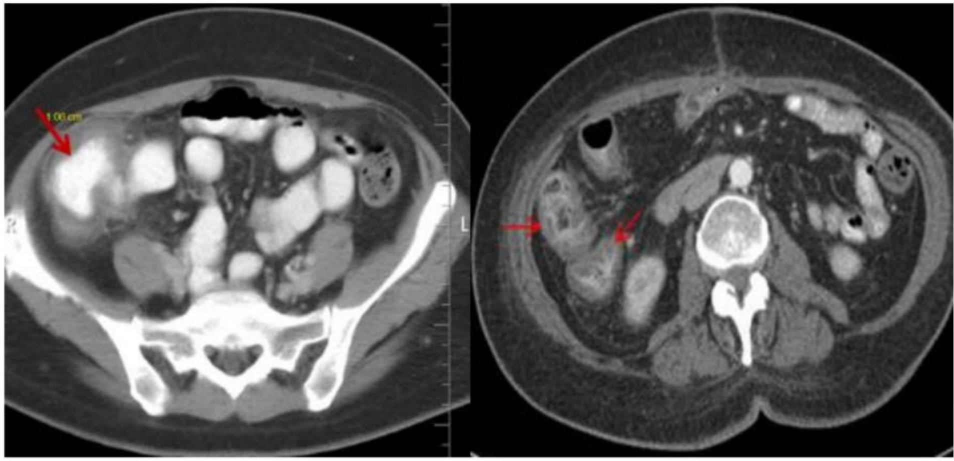
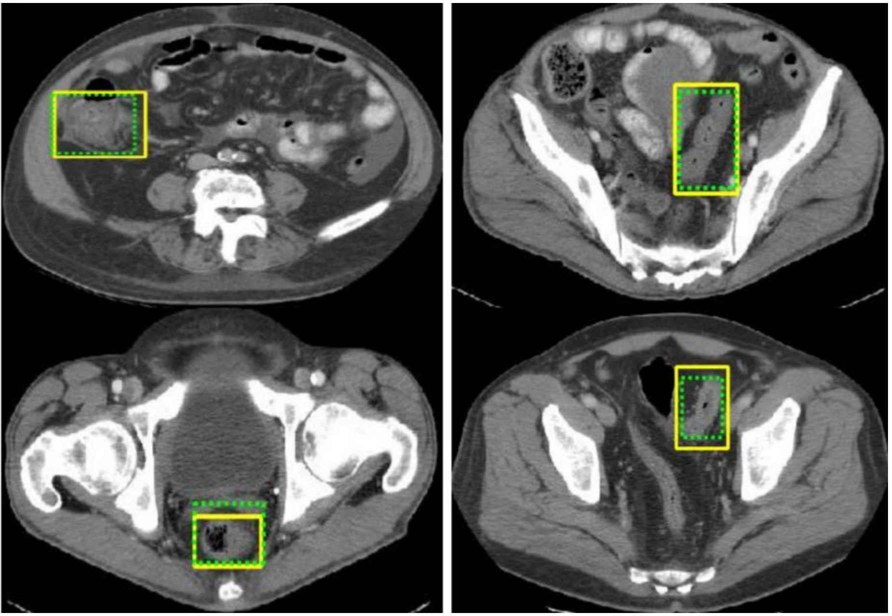


Figure 1. Examples of calcified plaques (red arrows) on abdominal (left) and pelvic (right) CT scans.



Atherosclerotic Vascular Calcification Detection and Segmentation on Low Dose Computed Tomography Scans …, Liu et al., IEEE ISBI 2017 Oral





***Detecting the undetectables?**

***Fitting in practical/real clinical settings in the wild??**

COLITIS DETECTION ON COMPUTED TOMOGRAPHY USING REGIONAL CONVOLUTIONAL NEURAL NETWORKS, Liu et al., IEEE ISBI 2016

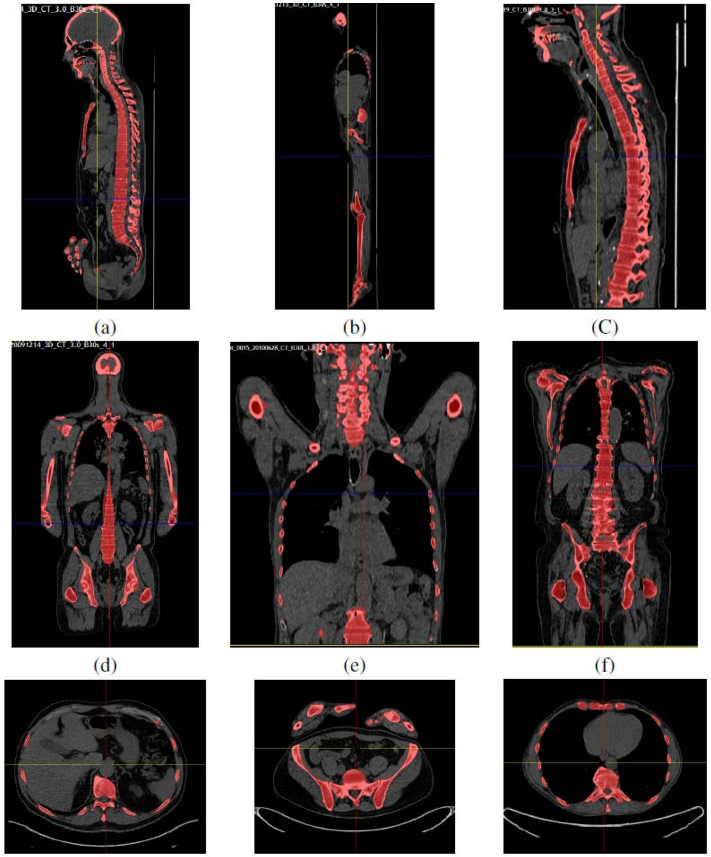
Three Key Problems (II)

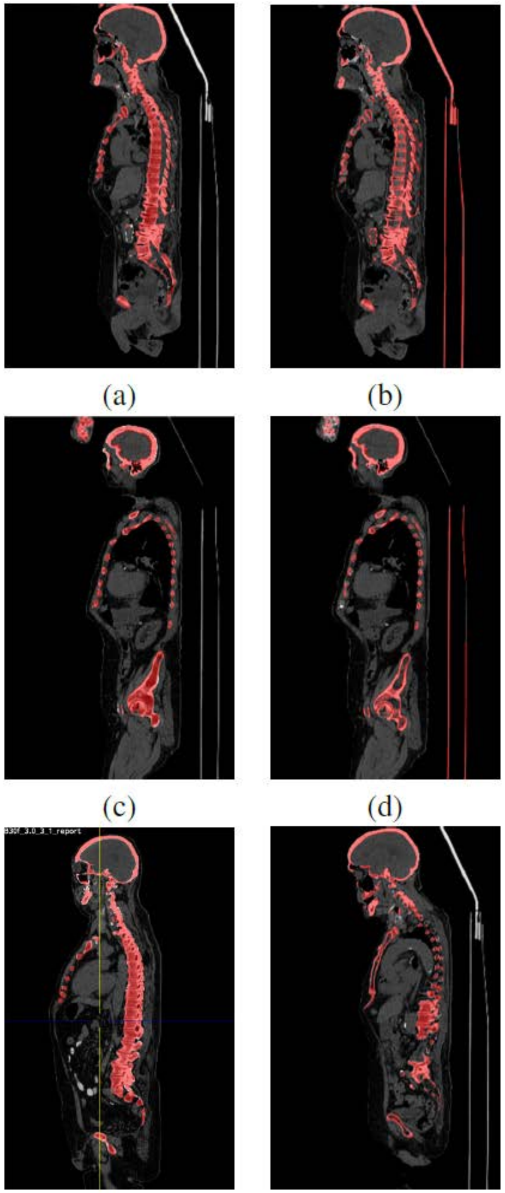
Semantic Segmentation in Medical Image Analysis

- "DeepOrgan" for **pancreas segmentation** (MICCAI 2015a) via scanning superpixels using multi-scale deep features ("Zoom-out") and probability map embedding.
- Deep segmentation on **pancreas** and **lymph node clusters** with Holistically-nested neural networks [Xie & Tu, 2015] as building blocks to learn unary (segmentation **mask**) and pairwise (labeling segmentation **boundary**) CRF terms + spatial aggregation or + structured optimization.
- The focus of three MICCAI 2016 papers since this is a much needed task → **Small** datasets; **(de-)compositional** representation is still the key. Scale up to thousands of patients if not more than that amount. Submissions to MICCAI 2017 → Effective and Efficient Precision Biomarkers, even predicting the future growth!

- **Clinical Impacts**: semantic segmentation can help compute clinically more accurate and desirable precision imaging bio-markers or measurements → precision imaging personalized treatment and therapy → less guess more doing …

Results on PET-CT Patient Datasets (pathological …)





Towards whole Body precision measurements or computable precision imaging biomarkers

→

“Robust Whole Body 3D Bone Masking via Bottom-up Appearance Modeling and Context Reasoning in Low-Dose CT Imaging”, Lu et al., IEEE WACV 2016

→

Bone Mineral Density (BMD) scores, Muscle/Fat volumetric measurements in whole body or arbitrary FOV imaging … lung nodules, bone lesions, head-and-neck radiation sensitive organs, segmenting flexible soft anatomical structures for precision medicine, all clinically needed!

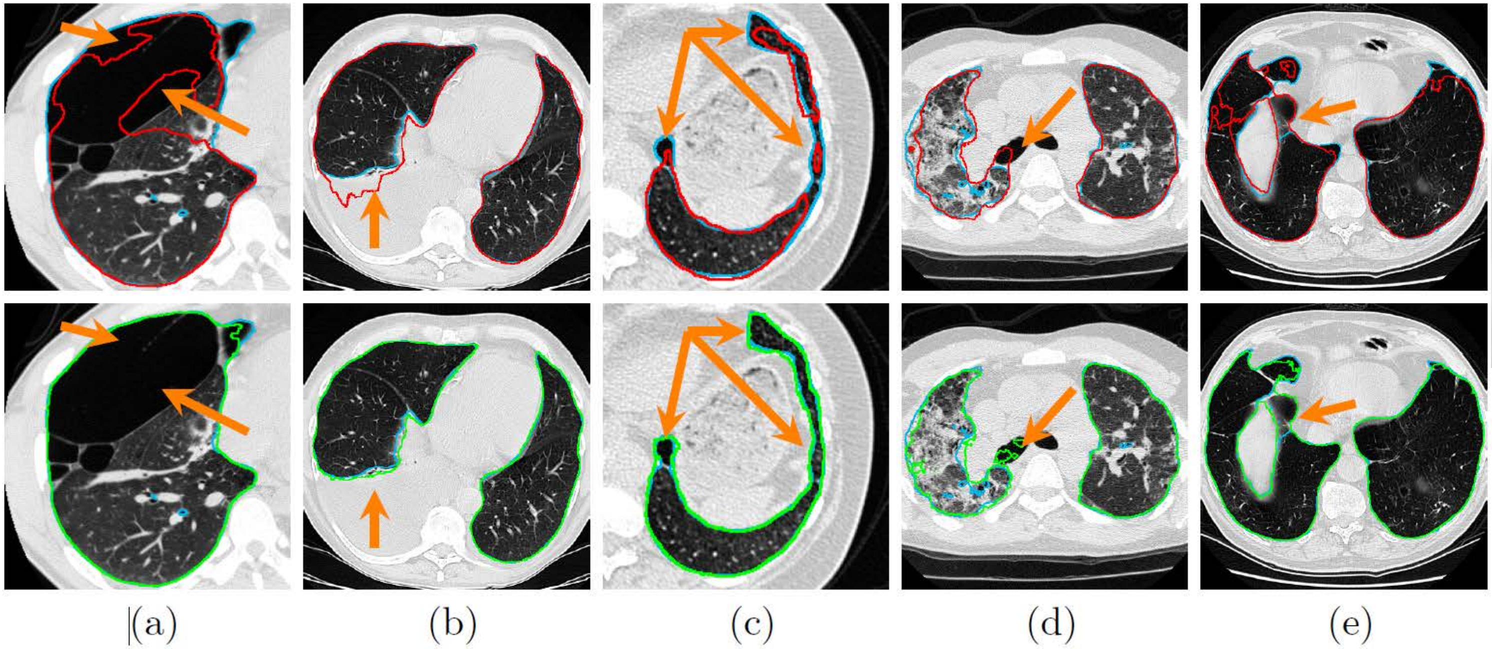
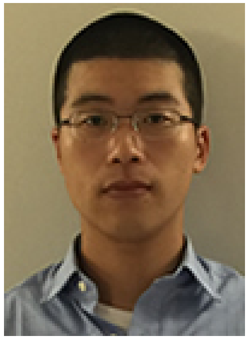


Fig. 2: Example masks of HNN and P-HNN, depicted in red and green, respectively. Ground truth masks are rendered in cyan. (a) HNN struggles to segment the pulmonary bullae, whereas P-HNN captures it. (b) Part of the pleural effusion is erroneously included by HNN, while left out of the P-HNN lung mask. (c) P-HNN is better able to capture finer details in the lung mask. (d) In this failure case, both HNN and P-HNN erroneously include the right main bronchus; however, P-HNN better captures infiltrate regions. (e) This erroneous ground-truth example, which was filtered out, fails to include a portion of the right lung. Both HNN and P-HNN capture the region, but P-HNN does a much better job of segmenting the rest of the lung.

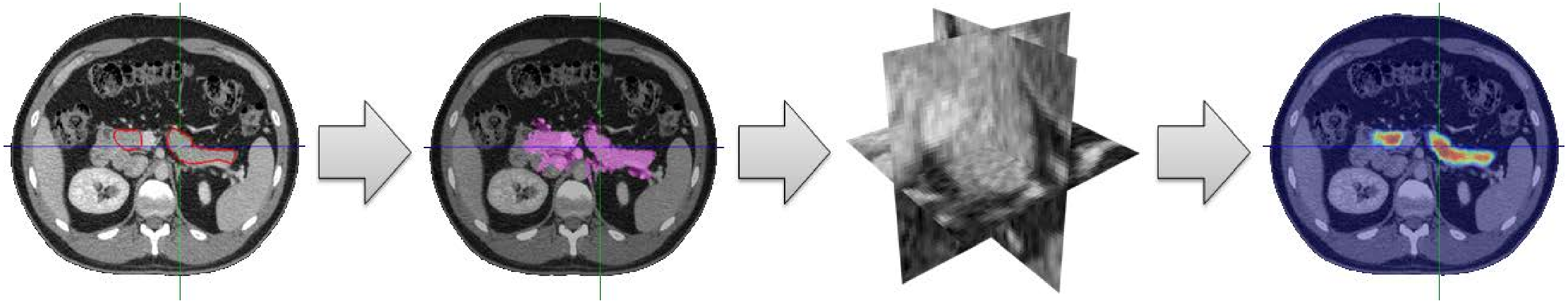
NSERC Fellow

A Roadmap of **Bottom-up Deep** Pancreas Segmentation: from Patch, Region, to Holistically-nested CNNs (HNN), P-HNN, Convolutional LSTM (context), …

Asst. Professor
Nagoya Uni.,
Japan

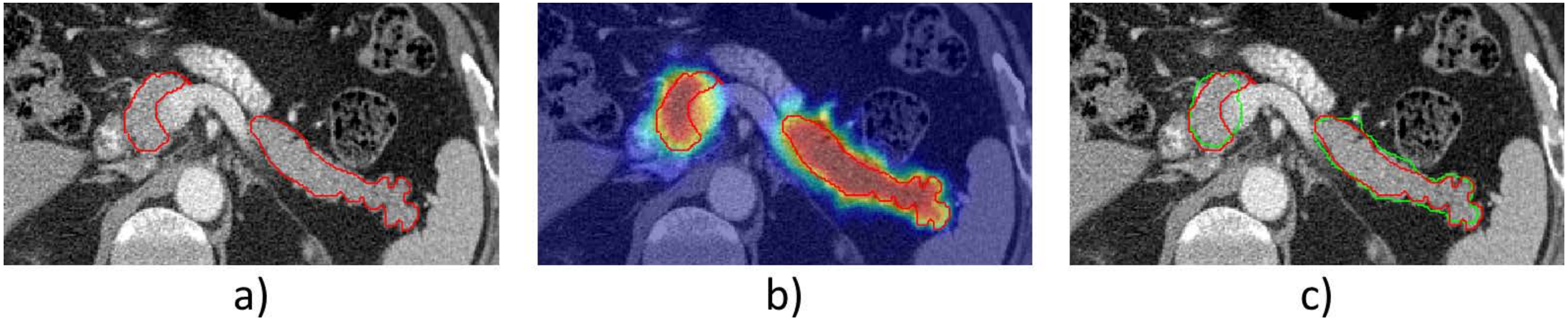


ISTP Fellow,
2012-2014



P-ConvNet

An Above-Average Example



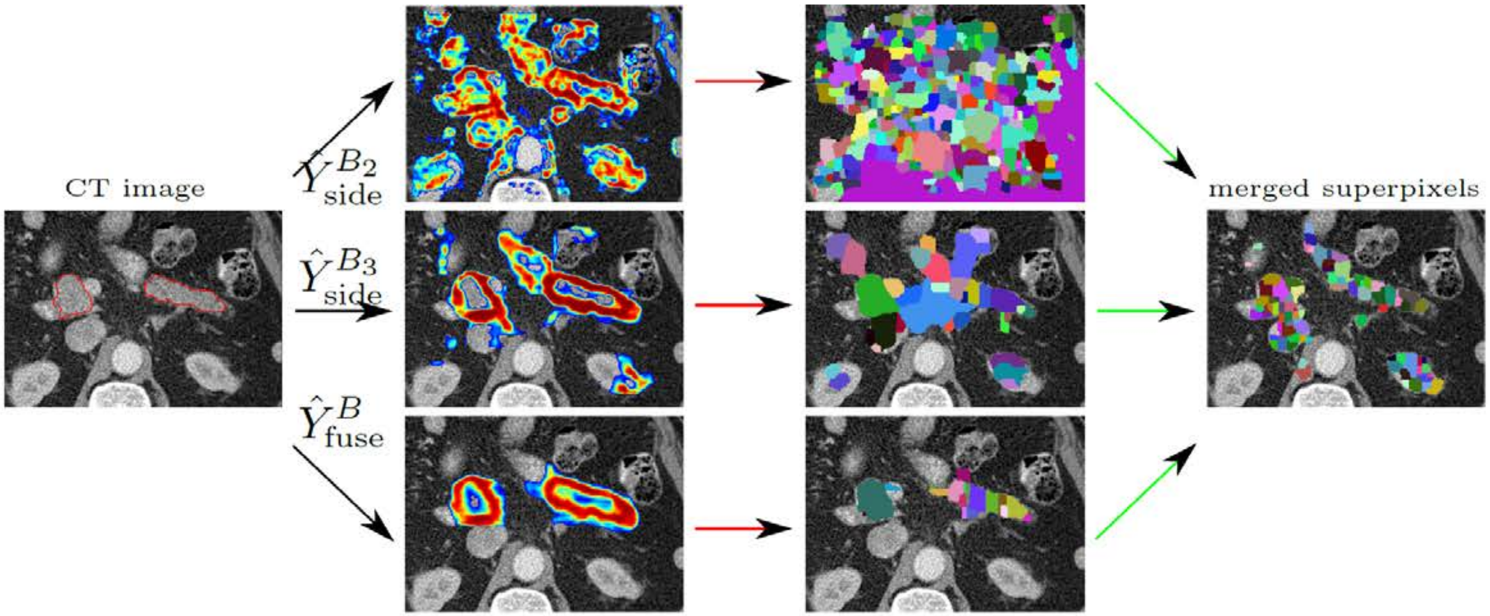

CT image
$\hat{Y}^{B_2}_{\text{side}}$
$\hat{Y}^{B_3}_{\text{side}}$
$\hat{Y}^{B}_{\text{fuse}}$
merged superpixels

Fig. 2: "Multiscale Combinatorial Grouping" (MCG) [16] on three different scales of learned boundary predication maps from **HNN-B**: $\hat{Y}^{B_2}_{\text{side}}$, $\hat{Y}^{B_3}_{\text{side}}$, and $\hat{Y}^{B}_{\text{fuse}}$ using the original CT image as input (shown with ground truth delineation of pancreas). MCG computes superpixels at each scale and produces a set of merged superpixel-based object proposals. We only visualize the boundary probabilities $p > 10\%$.

Improved pancreas segmentation accuracy over previous state-of-the-art work in Dice: from 68% to 84%; ASD: from 5~6mm to 0.7mm; computational time from 3 hours to >3 minutes!

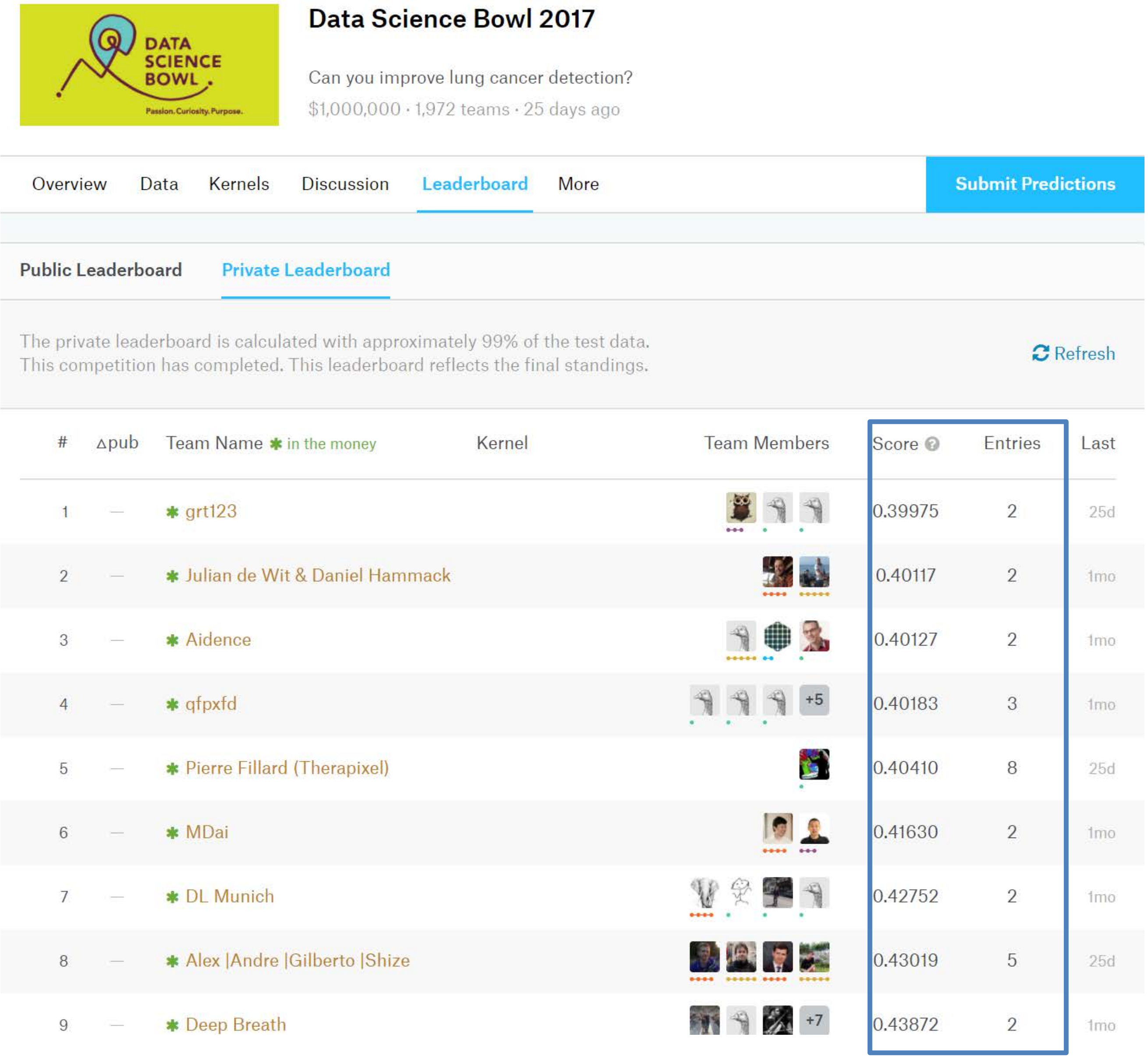
Three Key Problems (III)

Interleaved or Joint Text/Image Deep Mining on a Large-Scale Radiology Image Database → **“large”** datasets; **weak labels (~216K 2D key images/slices extracted from >60K unique patient studies)**

- Interleaved Text/Image Deep Mining on a Large-Scale Radiology Image Database (IEEE CVPR 2015, a proof of concept study)
- Interleaved Text/Image Deep Mining on a Large-Scale Radiology Image Database for Automated Image Interpretation (its extension, JMLR, 17(107):1−31, 2016)
- Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, (IEEE CVPR 2016)
- Unsupervised Category Discovery via Looped Deep Pseudo-Task Optimization Using a Large Scale Radiology Image Database, IEEE WACV 2017
- ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR 2017

➢ **Clinical Impacts**: eventually to build an automated mechanism to parse and learn from hospital scale PACS-RIS databases to derive semantics and knowledge … has to be ***deep learning*** based since effective image features are very hard to be hand-crafted cross different diseases, imaging protocols and modalities.

Q2: Are we at the edge of cracking radiology?



Data Science Bowl 2017

Can you improve lung cancer detection?

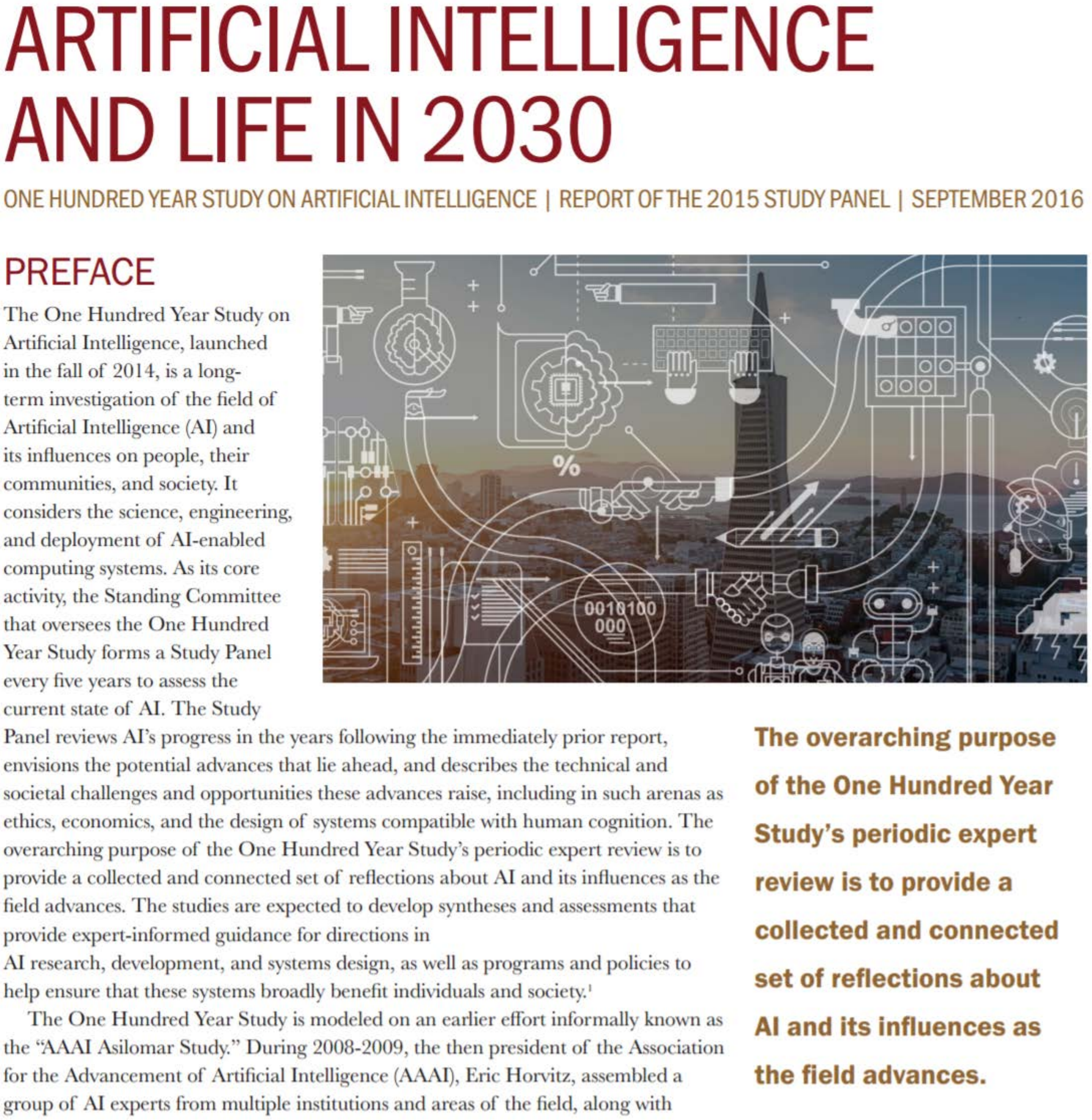
$1,000,000 · 1,972 teams · 25 days ago

The private leaderboard is calculated with approximately 99% of the test data.
This competition has completed. This leaderboard reflects the final standings.

#	∆pub	Team Name ✱ in the money	Kernel	Team Members	Score	Entries	Last
1	—	✱ grt123			0.39975	2	25d
2	—	✱ Julian de Wit & Daniel Hammack			0.40117	2	1mo
3	—	✱ Aidence			0.40127	2	1mo
4	—	✱ qfpxfd		+5	0.40183	3	1mo
5	—	✱ Pierre Fillard (Therapixel)			0.40410	8	25d
6	—	✱ MDai			0.41630	2	1mo
7	—	✱ DL Munich			0.42752	2	1mo
8	—	✱ Alex |Andre |Gilberto |Shize			0.43019	5	25d
9	—	✱ Deep Breath		+7	0.43872	2	1mo

***Issues/difficulties are beyond just datasets availability!**

** There are **many technical/methodological unknowns or challenges** to tackle in application performance requirements, problem setups, **label uncertainties** and more importantly, **proper image representations**, **Knowledge Ontology**, handling **long tail problems** gracefully without too embarrassing breakdown, etc …



ARTIFICIAL INTELLIGENCE AND LIFE IN 2030

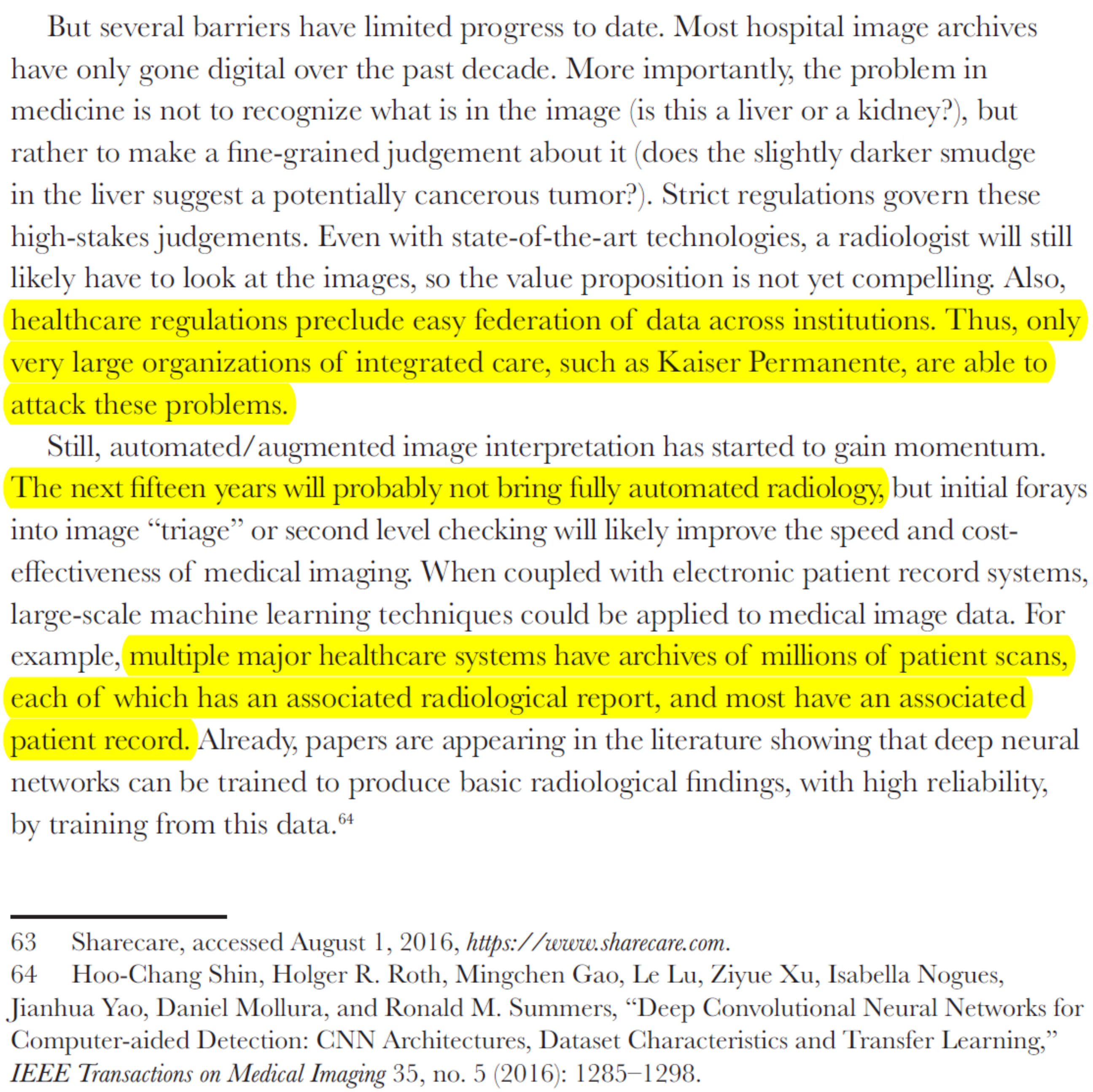
ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE | REPORT OF THE 2015 STUDY PANEL | SEPTEMBER 2016

PREFACE

The One Hundred Year Study on Artificial Intelligence, launched in the fall of 2014, is a long-term investigation of the field of Artificial Intelligence (AI) and its influences on people, their communities, and society. It considers the science, engineering, and deployment of AI-enabled computing systems. As its core activity, the Standing Committee that oversees the One Hundred Year Study forms a Study Panel every five years to assess the current state of AI. The Study Panel reviews AI's progress in the years following the immediately prior report, envisions the potential advances that lie ahead, and describes the technical and societal challenges and opportunities these advances raise, including in such arenas as ethics, economics, and the design of systems compatible with human cognition. The overarching purpose of the One Hundred Year Study's periodic expert review is to provide a collected and connected set of reflections about AI and its influences as the field advances. The studies are expected to develop syntheses and assessments that provide expert-informed guidance for directions in AI research, development, and systems design, as well as programs and policies to help ensure that these systems broadly benefit individuals and society.[1]

The One Hundred Year Study is modeled on an earlier effort informally known as the "AAAI Asilomar Study." During 2008-2009, the then president of the Association for the Advancement of Artificial Intelligence (AAAI), Eric Horvitz, assembled a group of AI experts from multiple institutions and areas of the field, along with

The overarching purpose of the One Hundred Year Study's periodic expert review is to provide a collected and connected set of reflections about AI and its influences as the field advances.



But several barriers have limited progress to date. Most hospital image archives have only gone digital over the past decade. More importantly, the problem in medicine is not to recognize what is in the image (is this a liver or a kidney?), but rather to make a fine-grained judgement about it (does the slightly darker smudge in the liver suggest a potentially cancerous tumor?). Strict regulations govern these high-stakes judgements. Even with state-of-the-art technologies, a radiologist will still likely have to look at the images, so the value proposition is not yet compelling. Also, healthcare regulations preclude easy federation of data across institutions. Thus, only very large organizations of integrated care, such as Kaiser Permanente, are able to attack these problems.

Still, automated/augmented image interpretation has started to gain momentum. The next fifteen years will probably not bring fully automated radiology, but initial forays into image "triage" or second level checking will likely improve the speed and cost-effectiveness of medical imaging. When coupled with electronic patient record systems, large-scale machine learning techniques could be applied to medical image data. For example, multiple major healthcare systems have archives of millions of patient scans, each of which has an associated radiological report, and most have an associated patient record. Already, papers are appearing in the literature showing that deep neural networks can be trained to produce basic radiological findings, with high reliability, by training from this data.[64]

[63] Sharecare, accessed August 1, 2016, *https://www.sharecare.com*.

[64] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers, "Deep Convolutional Neural Networks for Computer-aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging* 35, no. 5 (2016): 1285–1298.

Medical Dataset Availability is one of the Major Roadblocks and Helps are on the way!

- ➢ **Database #1: Interleaved or Joint Text/Image Deep Mining** on a Large-Scale Radiology Image Database → **"real PACS-large"** datasets; "**weak clinical annotations"**
  - Interleaved Text/Image Deep Mining on a Large-Scale Radiology Image Database, **IEEE CVPR 2015** (a proof of concept study)
  - Interleaved Text/Image Deep Mining on a Large-Scale Radiology Image Database for Automated Image Interpretation, **JMLR, 17(107):1−31, 2016**
  - Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition, **IEEE WACV, 2017**
  - **…**

✓ **Clinical Goal**: eventually to build an "**automated programmable mechanism"** to parse, extract and learn from **hospital-scale PACS-RIS databases,** to derive useful semantics and knowledge …

- ➢ ***Deep learning feature representation*** is a must since it is very hard to have effective hand-crafted image features cross different disease types, imaging protocols or modalities, if not at all impossible.
- ➢ **Algorithm innovations** to facilitate learning from "big data, weak label" large-scale retrospective clinical database!

NIH National Institutes of Health
Turning Discovery Into Health

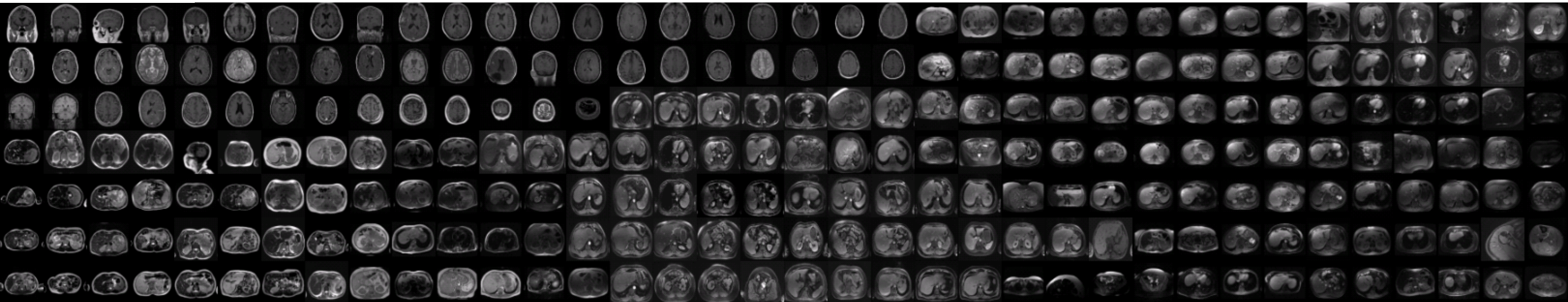
Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition

Xiaosong Wang, Le Lu, Hoo-chang Shin, Lauren Kim, Hadi Bagheri,
Isabella Nogues, Jianhua Yao and Ronald M. Summers

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory,
Department of Radiology and Imaging Sciences,
National Institutes of Health Clinical Center, Bethesda, MD 20892

US Patent Application, 62/302,096

WACV 2017



Motivation

- The availability of well-labeled data is the key for large scale machine learning, e.g., deep learning
- Labels for large medical imaging database are NOT available
- Conventional ways for collecting image labels are NOT applicable, e.g.
  - ❑ Google search followed by crowd-sourcing
  - ❑ Annotation on medical images requires professionals with clinical training

Large scale natural image datasets

VISUALGENOME
places
THE SCENE RECOGNITION DATABASE
IMAGENET
SUN
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning
COCO
Common Objects in Context

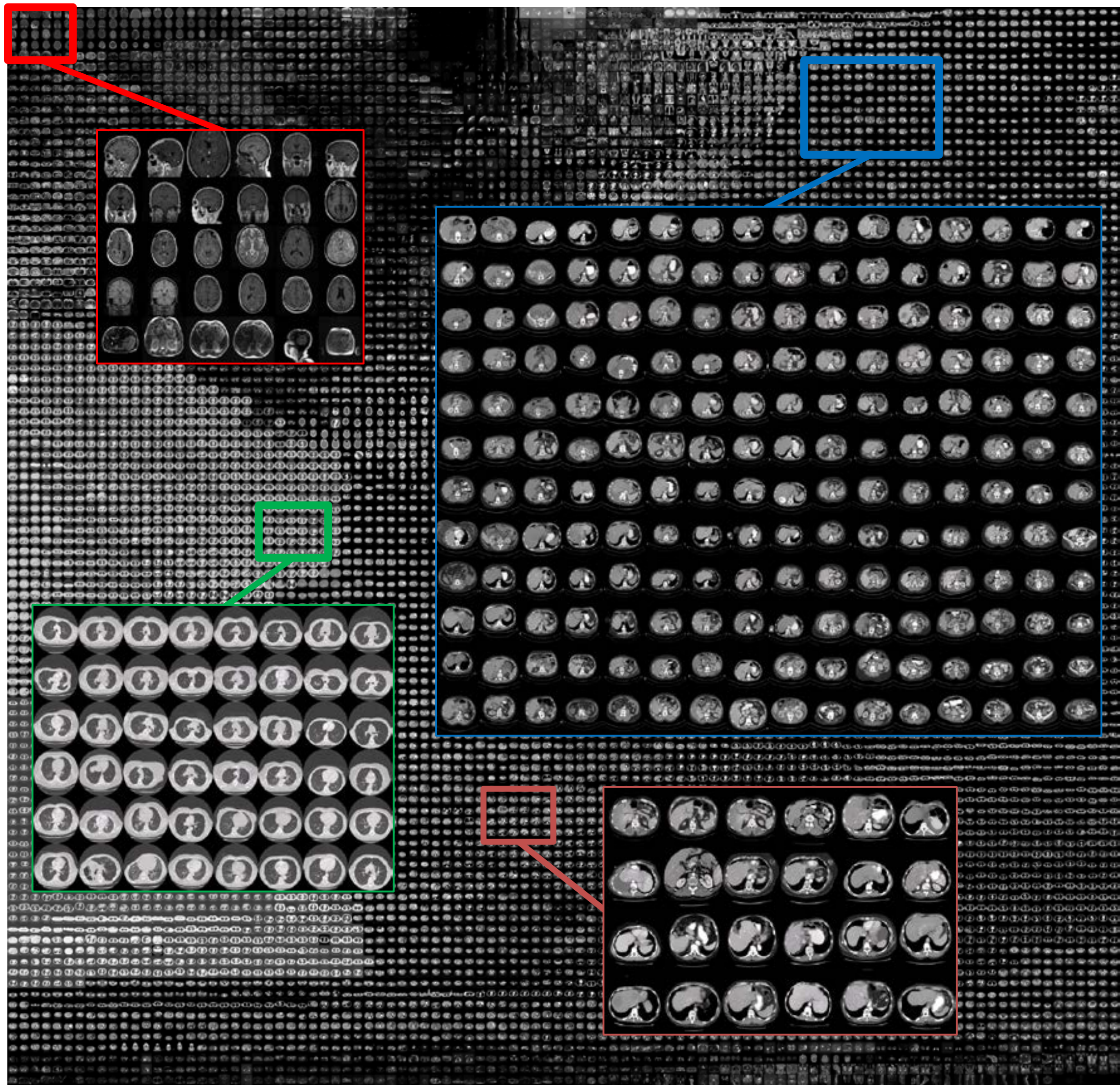
Large scale *<u>Medical Image</u>* *dataset*

?

* Dataset logos shown here are from respective public dataset websites.

Dataset

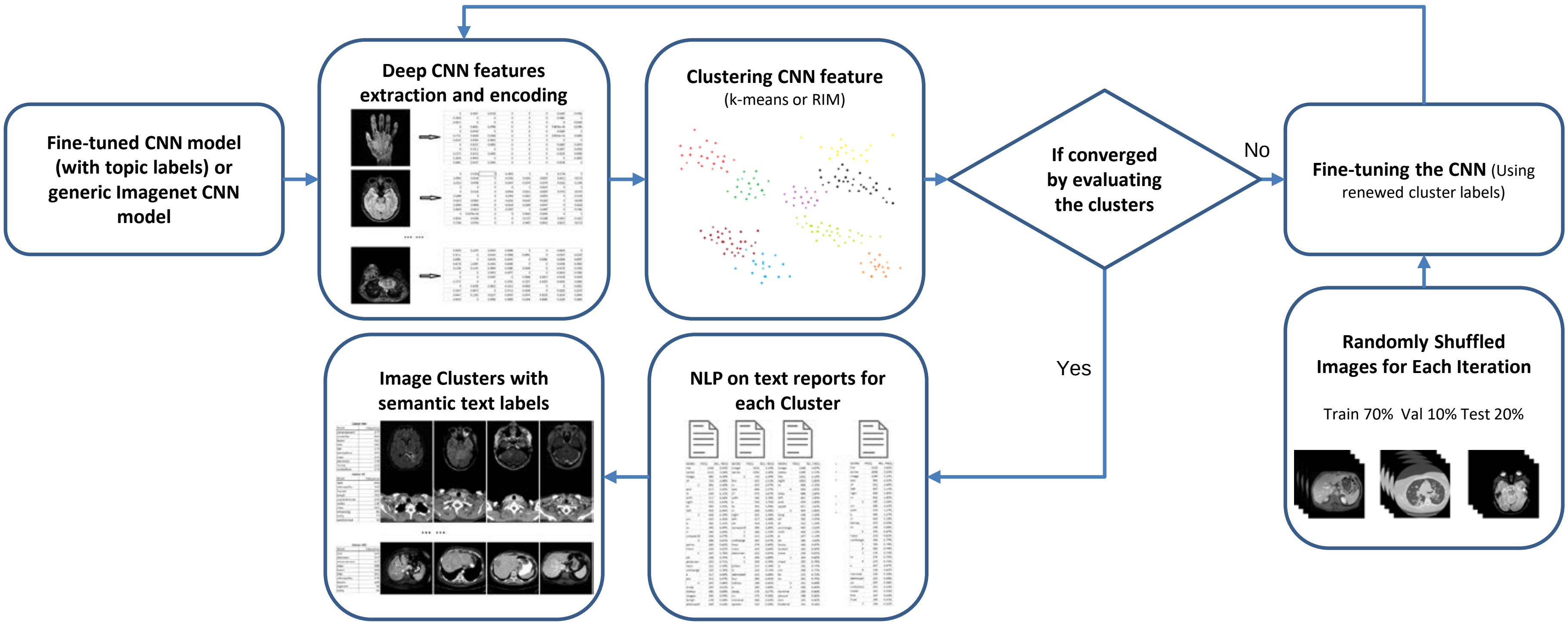
- A great treasure has been stored in our PACS system, i.e. images together with radiological reports.
- “Keyimage” dataset: 215,786 key images from 61,845 unique patient studies.
- Key images are significant one or more images in a study referenced in the linked radiological report.
- Key images are directly extracted from the DICOM file and resized as 256*256 bitmap images (.png).
- Their intensity ranges are rescaled using the default window settings stored in the DICOM header files.



* 10000 random images from the dataset, using CNN FC7 features of images embedded with t-SNE

Unsupervised Categorization

The proposed framework is designed towards automatic medical image annotation

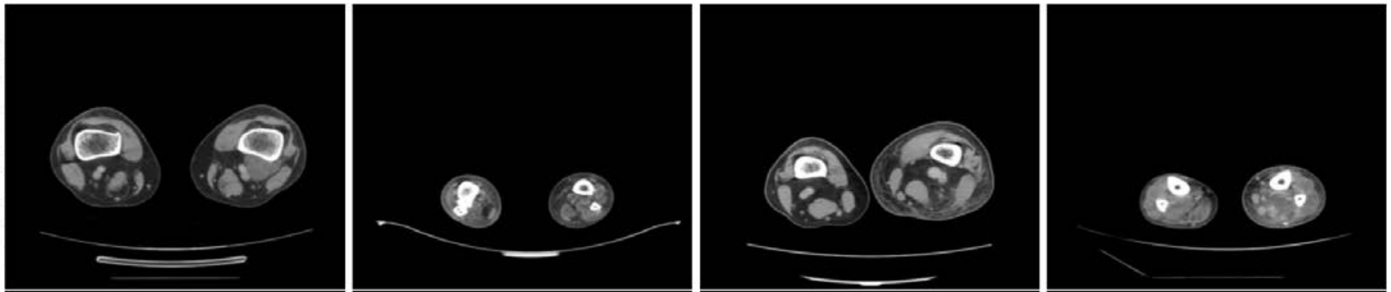

Fine-tuned CNN model (with topic labels) or generic Imagenet CNN model
Deep CNN features extraction and encoding
Clustering CNN feature (k-means or RIM)
If converged by evaluating the clusters
No
Fine-tuning the CNN (Using renewed cluster labels)
Randomly Shuffled Images for Each Iteration
Train 70% Val 10% Test 20%
Yes
NLP on text reports for each Cluster
Image Clusters with semantic text labels

- Hypothesized "convergence": better labels lead to better trained observable Convolutional Neural Network (CNN) models which consequently feed more effective deep image features to facilitate more meaningful clustering/labels.

Sample Categories

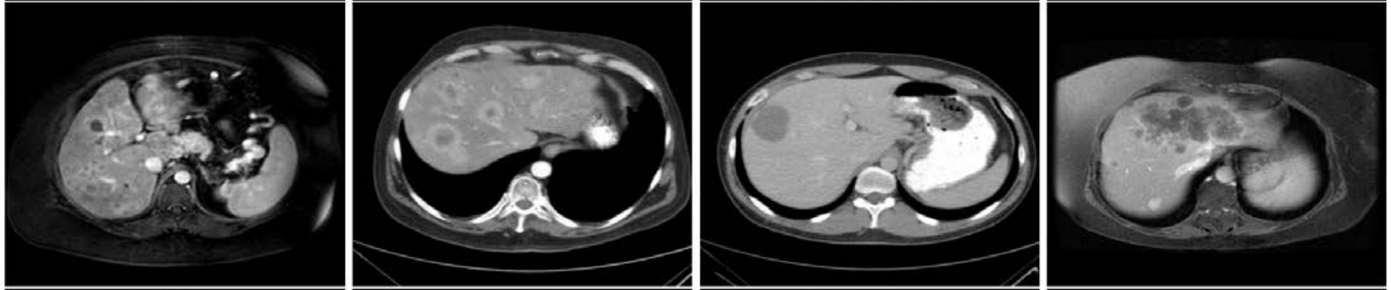
Cluster #14

Word	Frequency
calf	369
mass	263
subcutaneous	205
thigh	204
lesion	127
lower	124
enhancing	111
bone	105
fossa	92
nerve	88



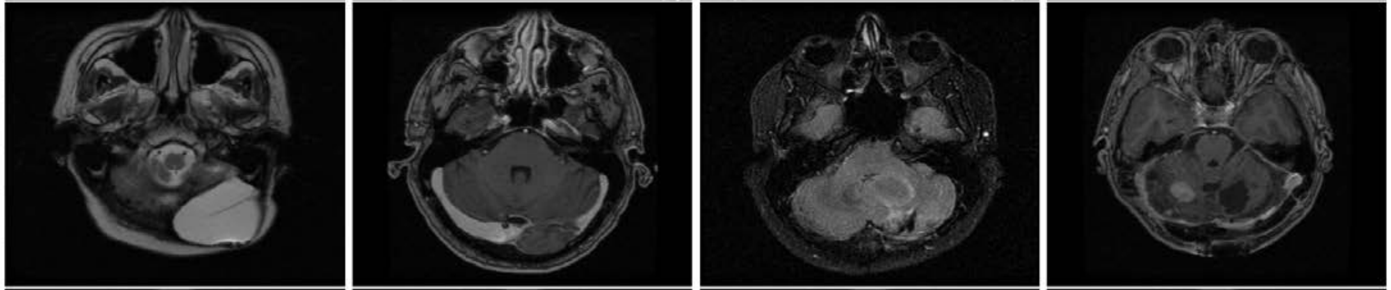
Cluster #23

Word	Frequency
liver	524
abdomen	337
enhancement	217
mass	198
lesion	168
lobe	161
adenopathy	119
lesions	109
segment	58
bulky	45



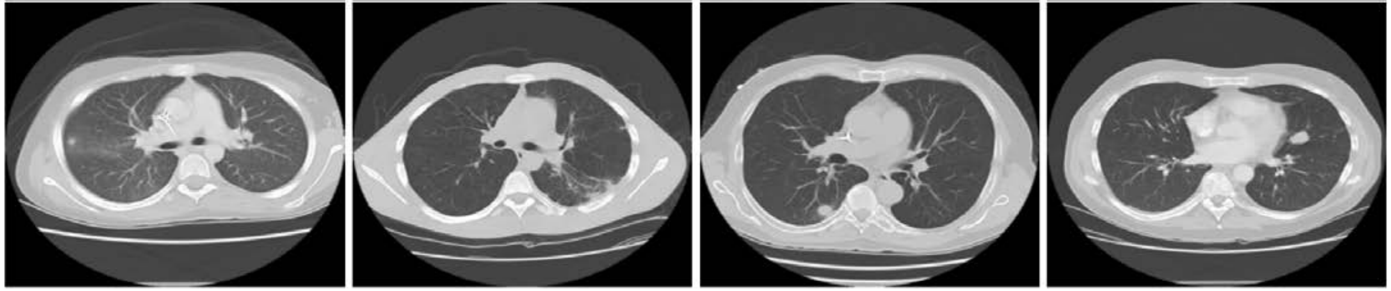
Cluster #64

Word	Frequency
enhancement	277
cerebellar	193
lesion	192
lobe	186
flair	173
hemisphere	155
mass	134
abnormal	119
frontal	115
cerebellum	113



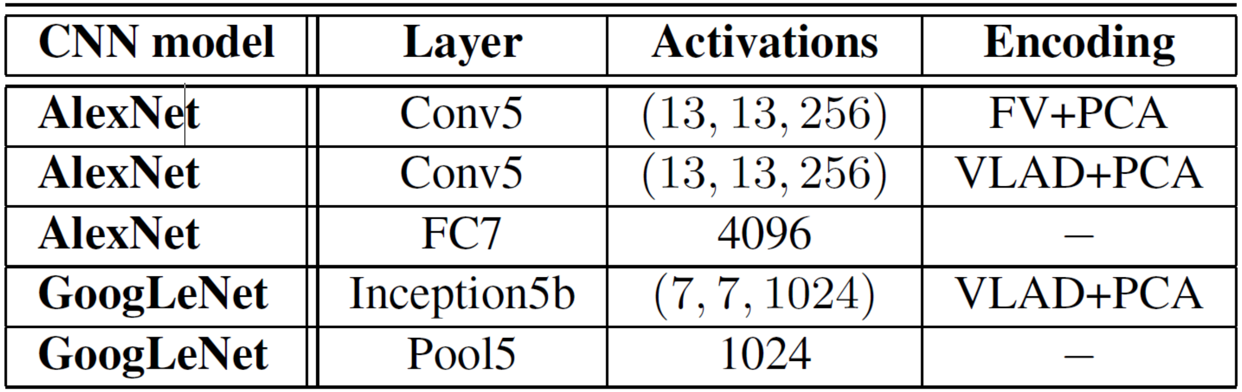
Cluster #224

Word	Frequency
lung	637
lobe	450
chest	361
mass	215
nodule	160
pleural	158
adenopathy	128
granulomata	111
atelectasis	86
pericardial	81



Experiment - CNN Setting

- The proposed framework is applicable to a variety of CNN models, by analyzing the CNN activations from layers of different depths.
- Encode the convolutional layer outputs in a form of dense pooling via Fisher Vector (FV) and Vector Locally Aggregated Descriptor (VLAD)
- Principal Component Analysis (PCA) is performed to reduce the dimensionality to 4096.

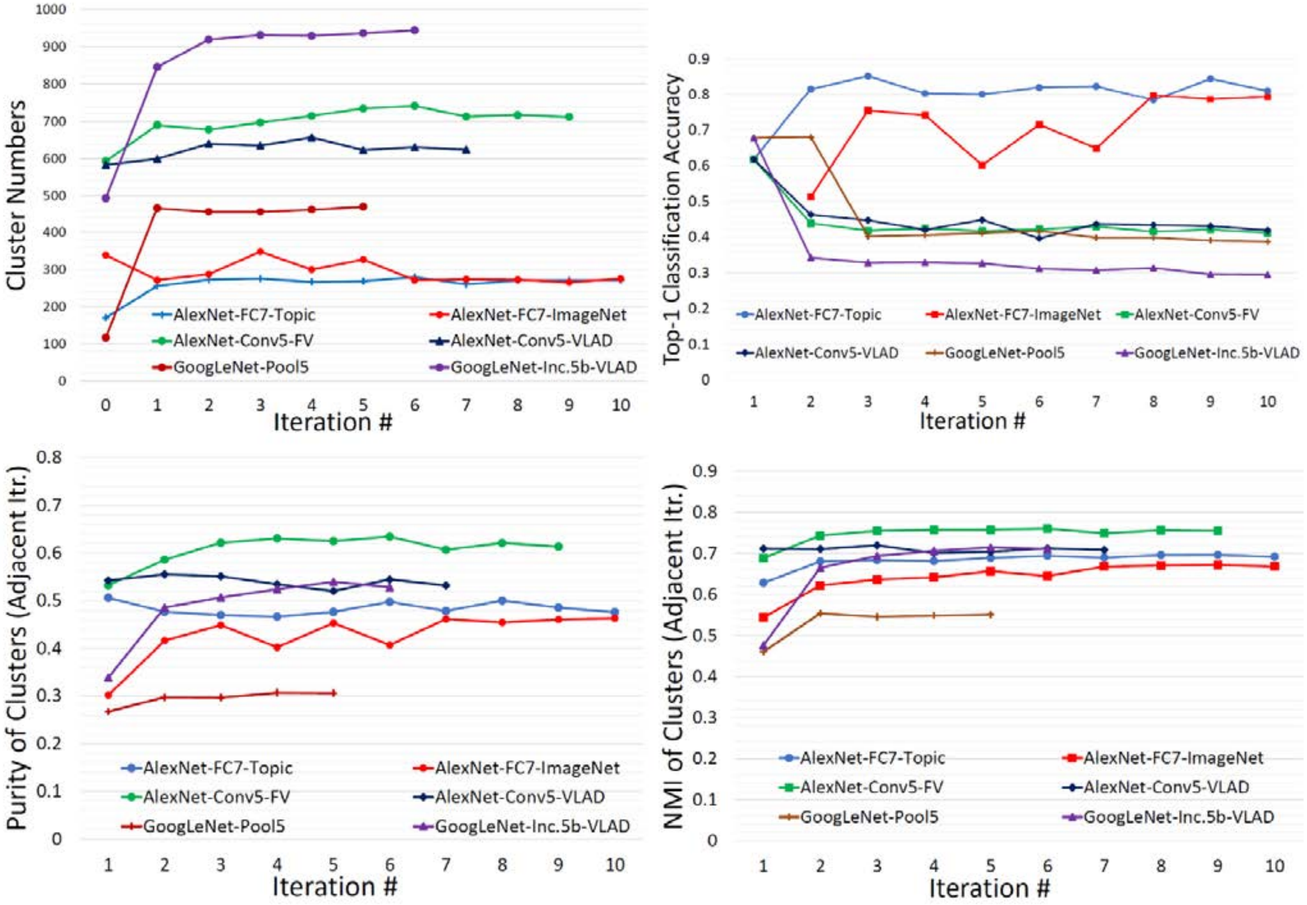


CNN model	Layer	Activations	Encoding
AlexNet	Conv5	$(13, 13, 256)$	FV+PCA
AlexNet	Conv5	$(13, 13, 256)$	VLAD+PCA
AlexNet	FC7	4096	–
GoogLeNet	Inception5b	$(7, 7, 1024)$	VLAD+PCA
GoogLeNet	Pool5	1024	–

Experiment - Convergence

- Clustering via K-means only or over-fragmented K-means followed by Regularized Information Maximization (as an effective model selection method), are extensively explored and empirically evaluated.
- Two convergence measurements have been adopted, i.e., Clustering Purity and Normalized Mutual Information (NMI).
- Newly generated clusters are better in terms of
  - ❑ Visually more coherent and discriminative from instances from other clusters
  - ❑ Balanced classes with approximately equivalent images per cluster
  - ❑ The number of clusters is self-adaptive according to the nature of data
  - ❑ Flexible to work with any clustering algorithm, no need to be differentiable or end-to-end trainable
  - ❑ Our conceptually simple unsupervised deep clustering algorithm shows effectiveness for other computer vision or medical imaging problems

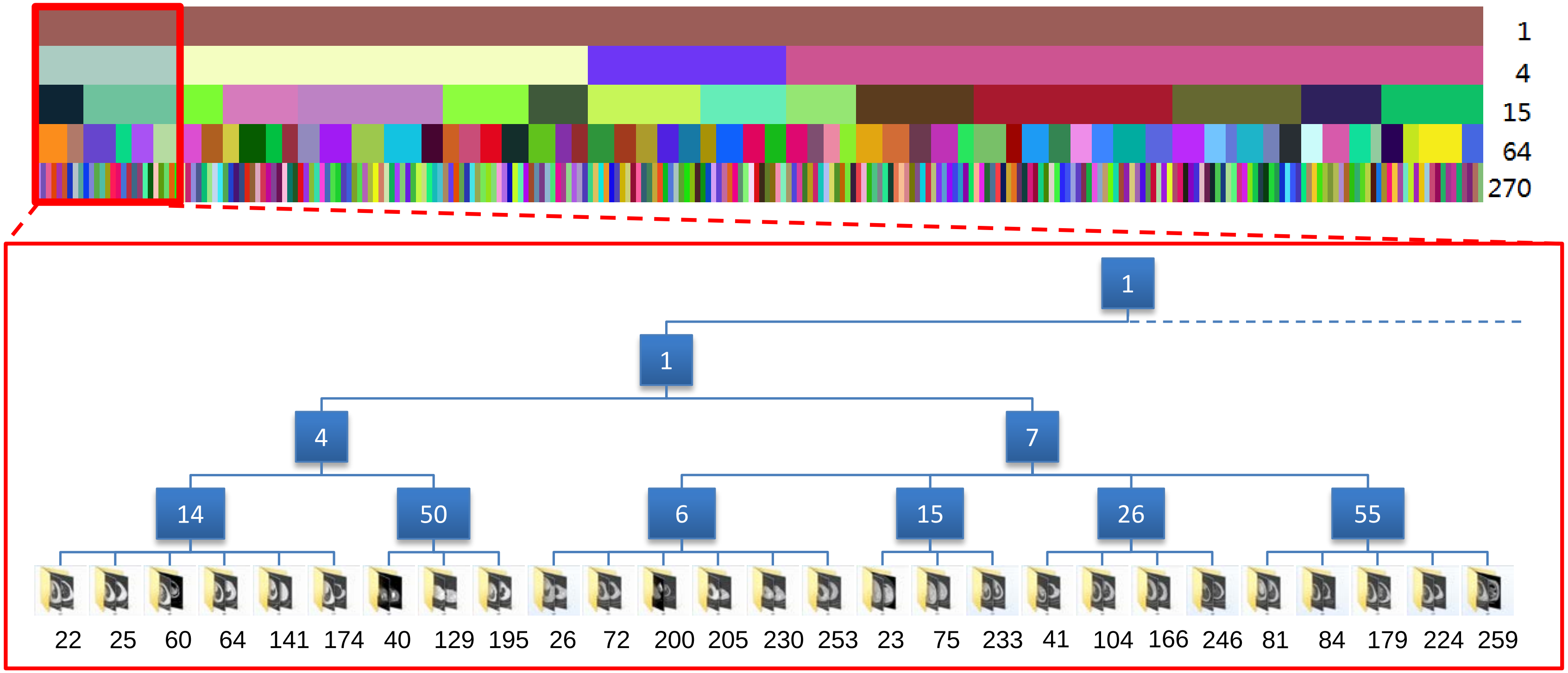
Quantitative Results



- The convergence of our categorization framework is measured and observed in the cluster-similarity measures, the CNN training classification accuracies and the self-adapted cluster number.
- AlexNet-FC7-Topic is preferred by two radiologists, which results in total 270 categories. The adopted FC7 feature is able to preserve the layout information of images.

Hierarchical Category Relationship

- Hierarchical category relationships in a tree-like structure can be naturally formulated and computed from the final pairwise CNN classification confusion measures. The resulting category tree has (270, 64, 15, 4, 1) different class labels from bottom (leaf) to top (root). The random color coded category tree is shown below.



Application on Scene Recognition

- Images from the same scene category may share similar object patches but are different in overall setting, e.g. buildings all have windows but in different style.
- Integrate patch mining as a form of image encoding into our LDPO framework and perform the categorization and patch mining iteratively.

MIT Indoor-67 (I-67)
indoor scenes | 67 classes
15620 images

Airport

Building-25 (B-25)
Architecture Style | 25 classes
4794 images

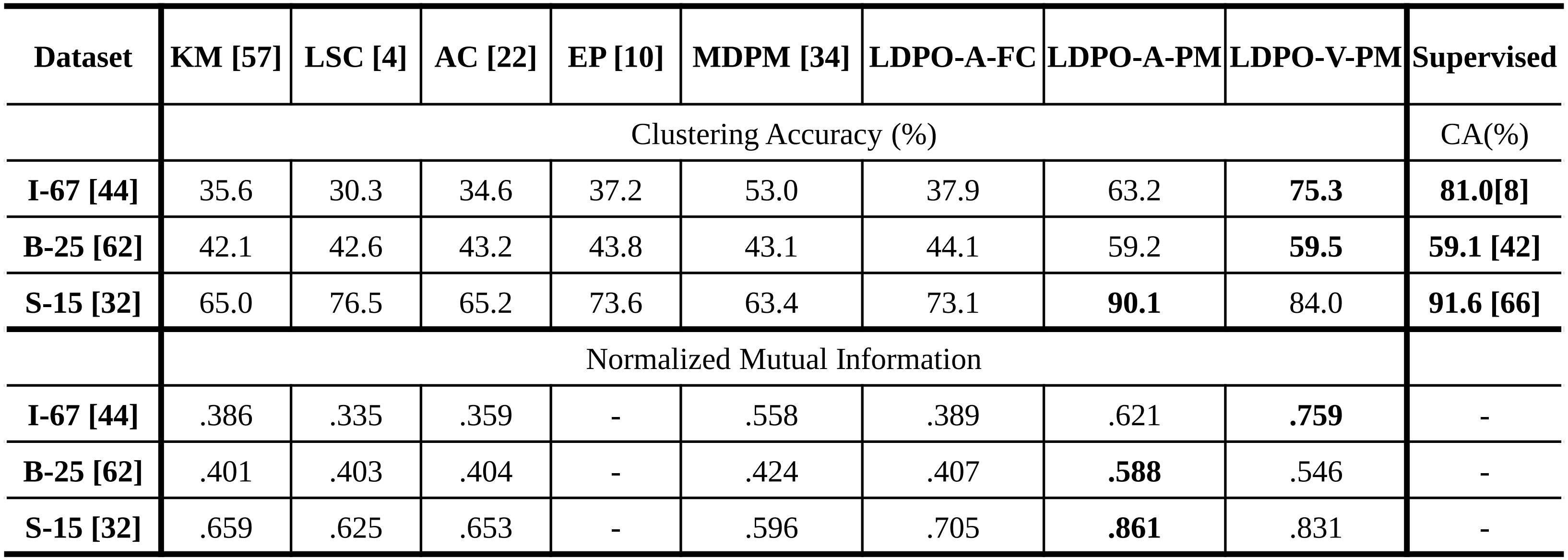
American Craftsman

Scene-15 (S-15)
Indoor & outdoor | 15 classes
4485 images

Bedroom

Evaluation on Clustering Accuracy

- The purity and NMI measurements are computed between the final LDPO clusters and GT scene classes (purity becomes the classification accuracy against GT).
- We compare the LDPO scene recognition performance to those of several popular clustering methods.
- The state-of-the-art fully-supervised scene Classification Accuracies (CA) for each dataset are also provided.



Dataset	KM [57]	LSC [4]	AC [22]	EP [10]	MDPM [34]	LDPO-A-FC	LDPO-A-PM	LDPO-V-PM	Supervised
	Clustering Accuracy (%)								CA(%)
I-67 [44]	35.6	30.3	34.6	37.2	53.0	37.9	63.2	**75.3**	**81.0[8]**
B-25 [62]	42.1	42.6	43.2	43.8	43.1	44.1	59.2	**59.5**	**59.1 [42]**
S-15 [32]	65.0	76.5	65.2	73.6	63.4	73.1	**90.1**	84.0	**91.6 [66]**
	Normalized Mutual Information								
I-67 [44]	.386	.335	.359	-	.558	.389	.621	**.759**	-
B-25 [62]	.401	.403	.404	-	.424	.407	**.588**	.546	-
S-15 [32]	.659	.625	.653	-	.596	.705	**.861**	.831	-

* KM: k-means; AC: agglomerative clustering ; LSC: large-scale spectral clustering ; EP: ensemble projection + k-means; MDPM: mid-level discriminative patch mining + k-means

Evaluation on Learned Image Features and Initialization Settings

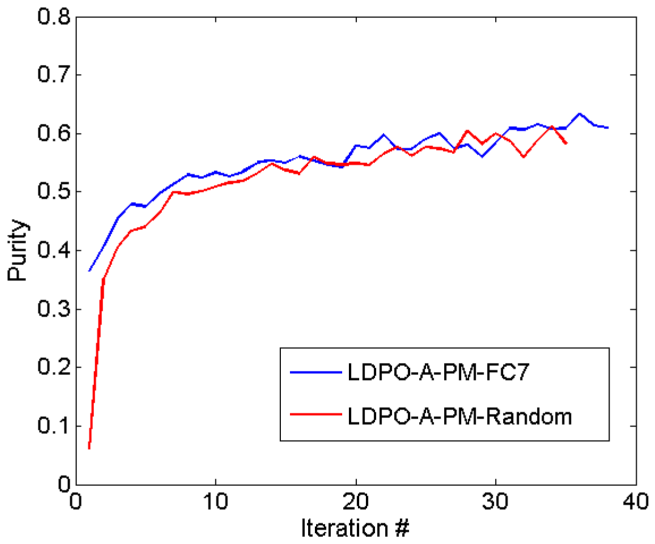
*** Learned image representation**:
1. Classification task on MIT-67, standard partition [44]
2. One-versus-all Liblinear classification on image features

× LDPO-V-PM-LL does not improve upon purely unsupervised LDPO-V-PM. This may indicate that LDPO-PM image representation is sufficient to separate images from different classes.

Method	Accuracy (%)
D-patch [53]	38.1
D-parts [54]	51.4
DMS [13]	64.0
MDPM-Alex [34]	64.1
MDPM-VGG [34]	77.0
MetaObject [60]	78.9
FC (VGG)	68.87
CONV-FV (VGG) [8]	**81.0**
LDPO-V-PM-LL	72.5
LDPO-V-PM	**75.3**

*** Different initialization settings**:
1. Random initialization
2. Image labels obtained from k-means clustering on FC7 features of an ImageNet pretrained AlexNet

✓ Both schemes ultimately converge to similar performance levels and it suggests that LDPO convergence is insensitive to the chosen initialization.



*Both results are computed using MIT Indoor-67 dataset.

Conclusion

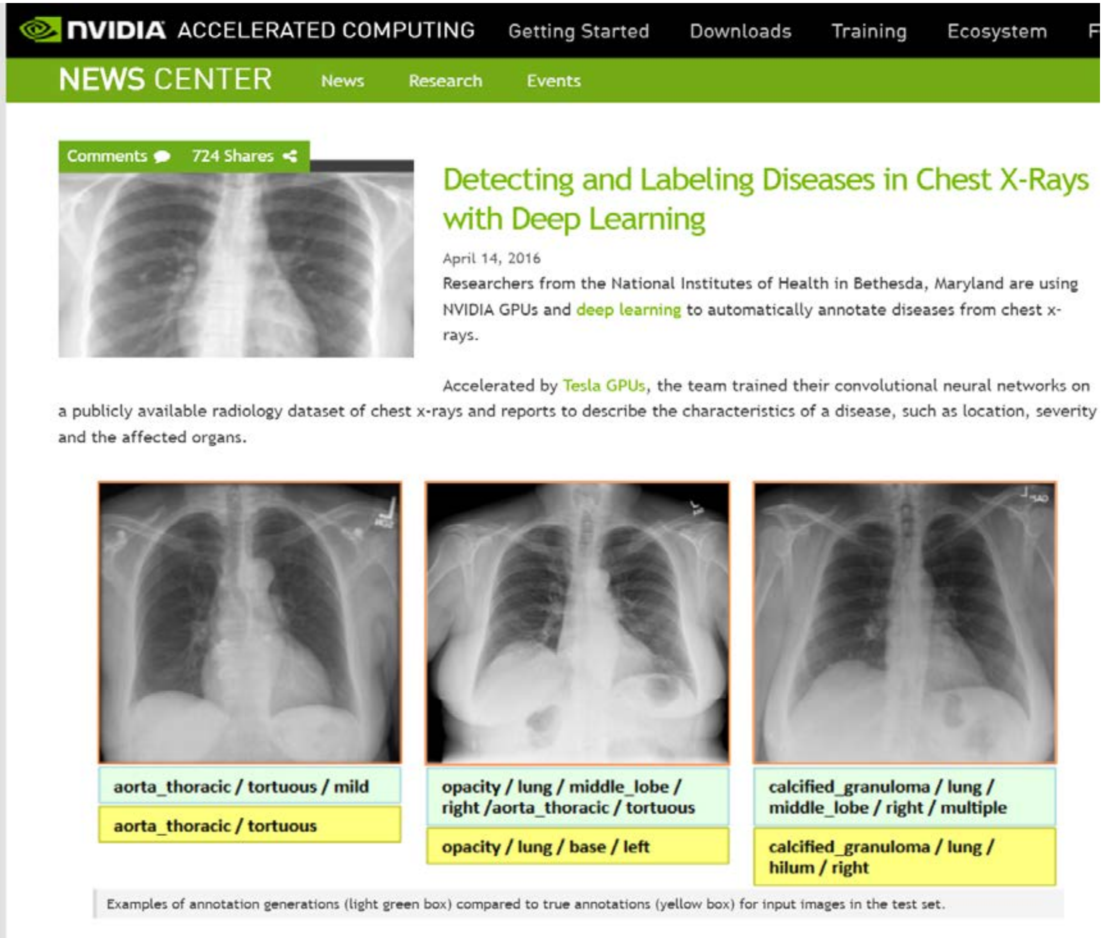
- Now It is time to **wake up the huge collection of clinical data sleeping in the PACS** and put it to work!
- A novel looped deep pseudo-task optimization framework is presented for category discovery from a large-scale medical image database.
- Extracted categories are visually more coherent and semantically meaningful (manually verified by experienced radiologists)
- We systematically and extensively conduct experiments under different settings of the proposed framework to validate and evaluate its quantitative and qualitative performance on two different types of dataset → effectively applicable to other CAD problems by exploiting finer-grained category information in an unsupervised manner.
- The measurable "convergence" makes the ill-posed auto-annotation problem well constrained, at no human labeling cost → towards building **radiology (anatomical & pathological) ontology** image database!

Learning to Read Chest X-ray using Deep Neural Networks (a little more like humans' interpretation?)

[Shin et al., IEEE CVPR 2016, US Patent Application: 62/302,084]

Lung diseases killing 4 million people every year, in comparison to Nearly **1.3 million people** die in **road crashes** each year!

Statistics from internet …



Detecting and Labeling Diseases in Chest X-Rays with Deep Learning

April 14, 2016

Researchers from the National Institutes of Health in Bethesda, Maryland are using NVIDIA GPUs and deep learning to automatically annotate diseases from chest x-rays.

Accelerated by Tesla GPUs, the team trained their convolutional neural networks on a publicly available radiology dataset of chest x-rays and reports to describe the characteristics of a disease, such as location, severity and the affected organs.

aorta_thoracic / tortuous / mild

aorta_thoracic / tortuous

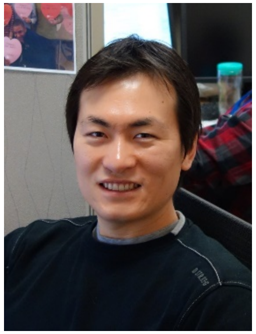
opacity / lung / middle_lobe / right /aorta_thoracic / tortuous

opacity / lung / base / left

calcified_granuloma / lung / middle_lobe / right / multiple

calcified_granuloma / lung / hilum / right

Examples of annotation generations (light green box) compared to true annotations (yellow box) for input images in the test set.



National Institutes of Health
Turning Discovery Into Health

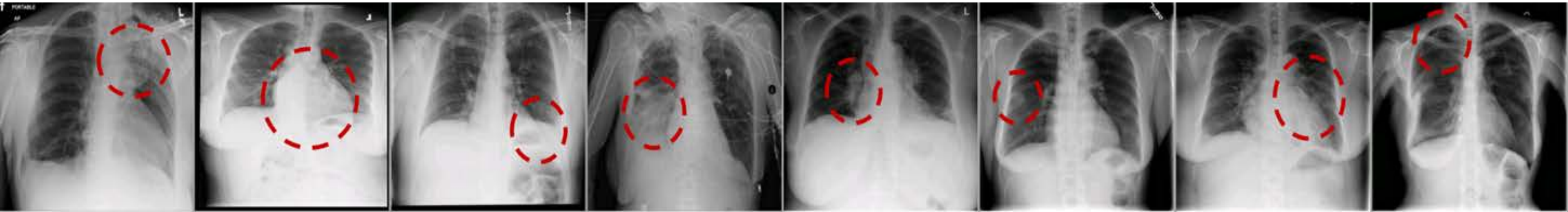
ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases

Xiaosong Wang[1] , Yifan Peng[2], Le Lu[1], Zhiyong Lu[2], Mohammadhadi Bagheri[1], Ronald M. Summers[1]

1 Department of Radiology and Imaging Sciences, Clinical Center
2 National Center for Biotechnology Information, National Library of Medicine
National Institutes of Health, Bethesda, MD 20892

US Patent Application, 62/476,029

CVPR July 21-26 2017 HONOLULU

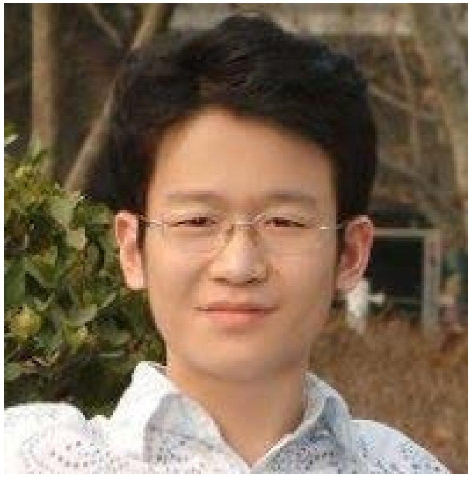


Motivation

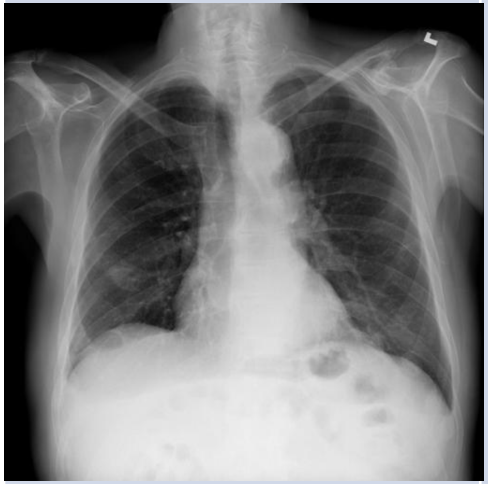
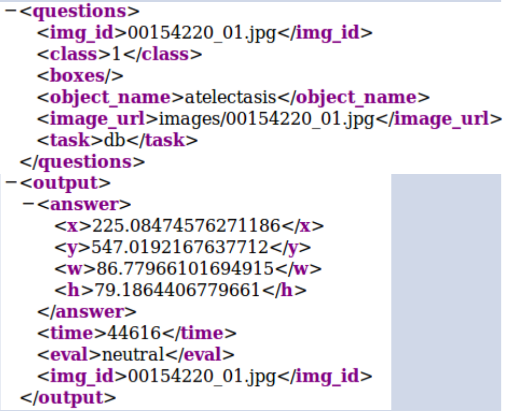
1. Build a large-scale Chest X-rays dataset to facilitate the data-hungry deep learning paradigms
2. Critical/common disease patterns (predefined by radiologist)
3. Multiple labels (disease patterns) for images
4. Radiological report for each image
5. Small number of bounding boxes (outline the location of disease symptoms) for each disease category
6. Potential applications:
   - A. Disease detection/classification
   - B. Disease localization
   - C. Automatic radiological report generation
   - …

Dataset 2#: A Hospital-scale Chest X-ray Dataset

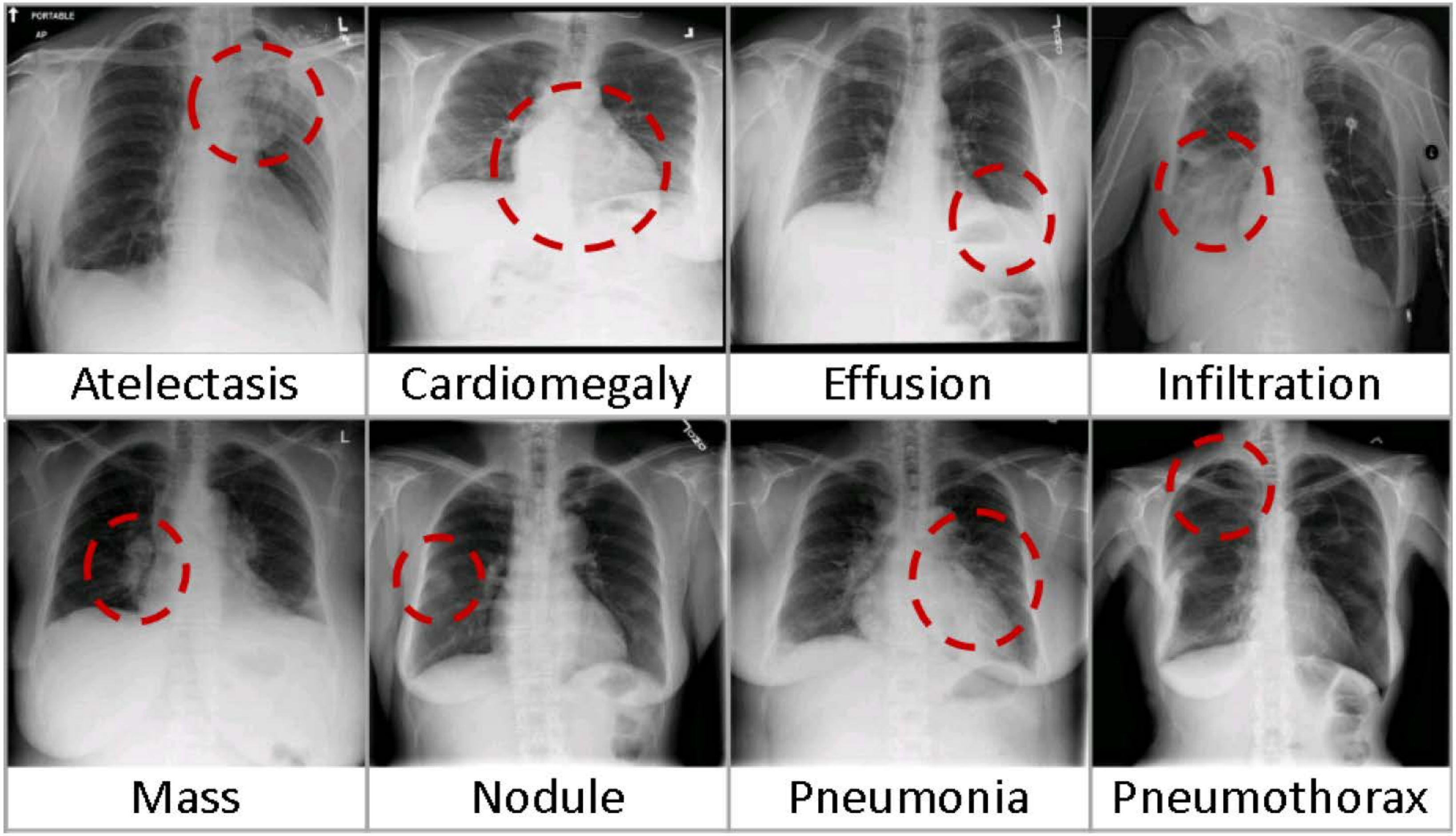
(Joint efforts by NIH-CC and NIH-NLM)



A Sample Entry

Image	Report	Label	Bounding Box (optional)
	findings: pa and lateral views of the chest demonstrate significantly improved bilateral lower lung field interstitial markings compatible with linear atelectasis. unchanged right 9th rib fracture peripherally. unchanged ossification left coracoacromial ligament. the cardiac and mediastinal contours are stable. impression: improved bilateral lower lung field linear atelectasis.	atelectasis	 −<questions> <img_id>00154220_01.jpg</img_id> <class>1</class> <boxes/> <object_name>atelectasis</object_name> <image_url>images/00154220_01.jpg</image_url> <task>db</task> </questions> −<output> −<answer> <x>225.08474576271186</x> <y>547.0192167637712</y> <w>86.77966101694915</w> <h>79.1864406779661</h> </answer> <time>44616</time> <eval>neutral</eval> <img_id>00154220_01.jpg</img_id> </output>

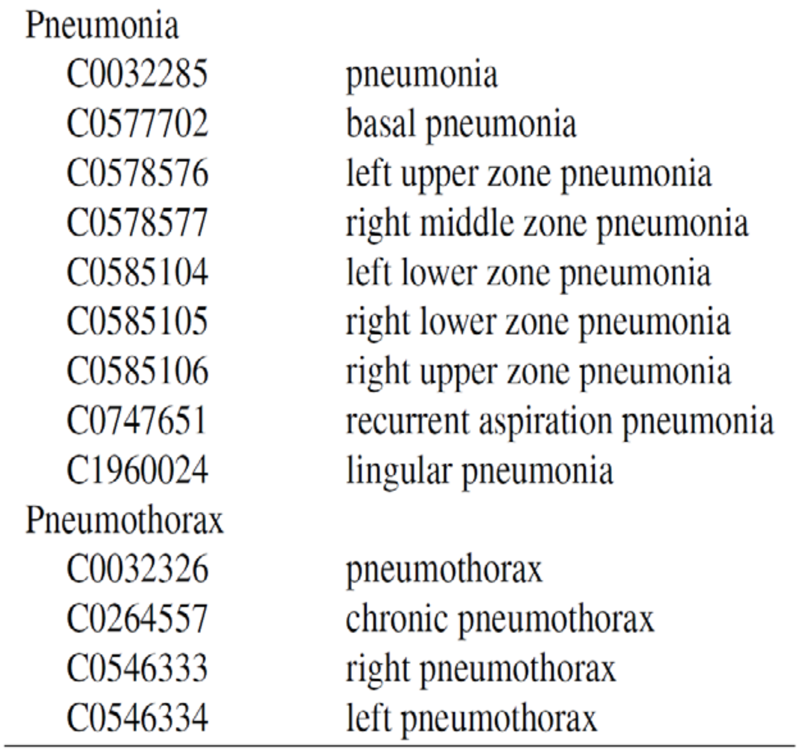
8 Common Thorax Diseases



Two-stage NLP of Radiology Reports

Stage 1: Pathology Detection

- **DNorm** is used to map every mention of keywords in a report to a unique concept ID in the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), a standardized vocabulary of clinical terminology for the electronic exchange of clinical health information.
- Another ontology-based approach, **MetaMap**, is adopted for the detection of Unified Medical Language System (UMLS) Metathesaurus.
- The results of DNorm and MetaMap are merged



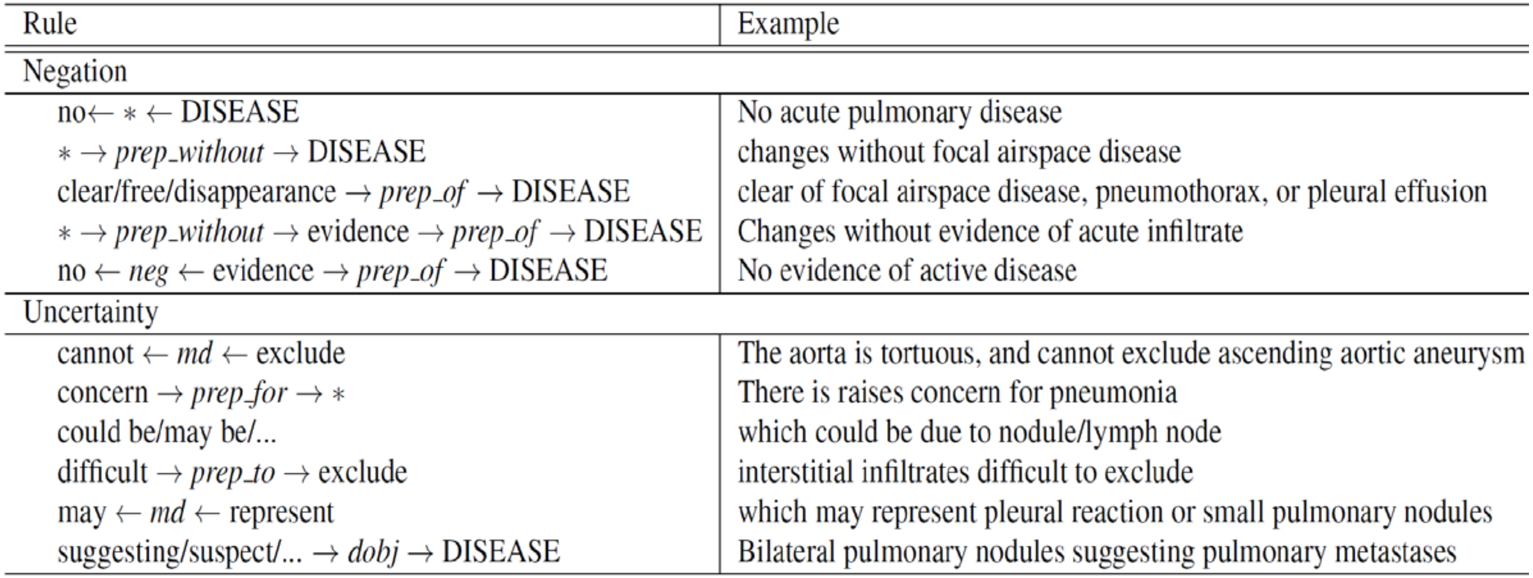
Pneumonia		
	C0032285	pneumonia
	C0577702	basal pneumonia
	C0578576	left upper zone pneumonia
	C0578577	right middle zone pneumonia
	C0585104	left lower zone pneumonia
	C0585105	right lower zone pneumonia
	C0585106	right upper zone pneumonia
	C0747651	recurrent aspiration pneumonia
	C1960024	lingular pneumonia
Pneumothorax		
	C0032326	pneumothorax
	C0264557	chronic pneumothorax
	C0546333	right pneumothorax
	C0546334	left pneumothorax

Sample SNOMED-CT concepts

Two-stage NLP of Radiology Reports

Stage 2: Removal of negation and uncertainty

- Rule out those negated pathological statements and uncertain mentions of findings and diseases
- Defined the rules on the dependency graph, by utilizing the dependency label and direction information between words, e.g.,



Rule	Example
Negation	
no← ∗ ← DISEASE	No acute pulmonary disease
∗ → *prep_without* → DISEASE	changes without focal airspace disease
clear/free/disappearance → *prep_of* → DISEASE	clear of focal airspace disease, pneumothorax, or pleural effusion
∗ → *prep_without* → evidence → *prep_of* → DISEASE	Changes without evidence of acute infiltrate
no ← *neg* ← evidence → *prep_of* → DISEASE	No evidence of active disease
Uncertainty	
cannot ← *md* ← exclude	The aorta is tortuous, and cannot exclude ascending aortic aneurysm
concern → *prep_for* → ∗	There is raises concern for pneumonia
could be/may be/...	which could be due to nodule/lymph node
difficult → *prep_to* → exclude	interstitial infiltrates difficult to exclude
may ← *md* ← represent	which may represent pleural reaction or small pulmonary nodules
suggesting/suspect/... → *dobj* → DISEASE	Bilateral pulmonary nodules suggesting pulmonary metastases

Two-stage NLP of Radiology Reports

Stage 2: Removal of negation and uncertainty

- Rule out those negated pathological statements and uncertain mentions of findings and diseases
- Defined the rules on the dependency graph, by utilizing the dependency label and direction information between words, e.g.,

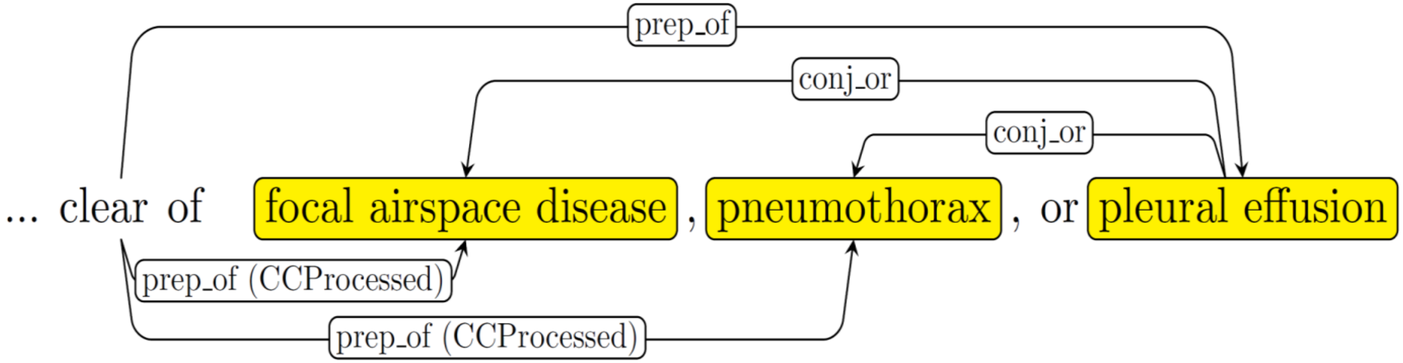
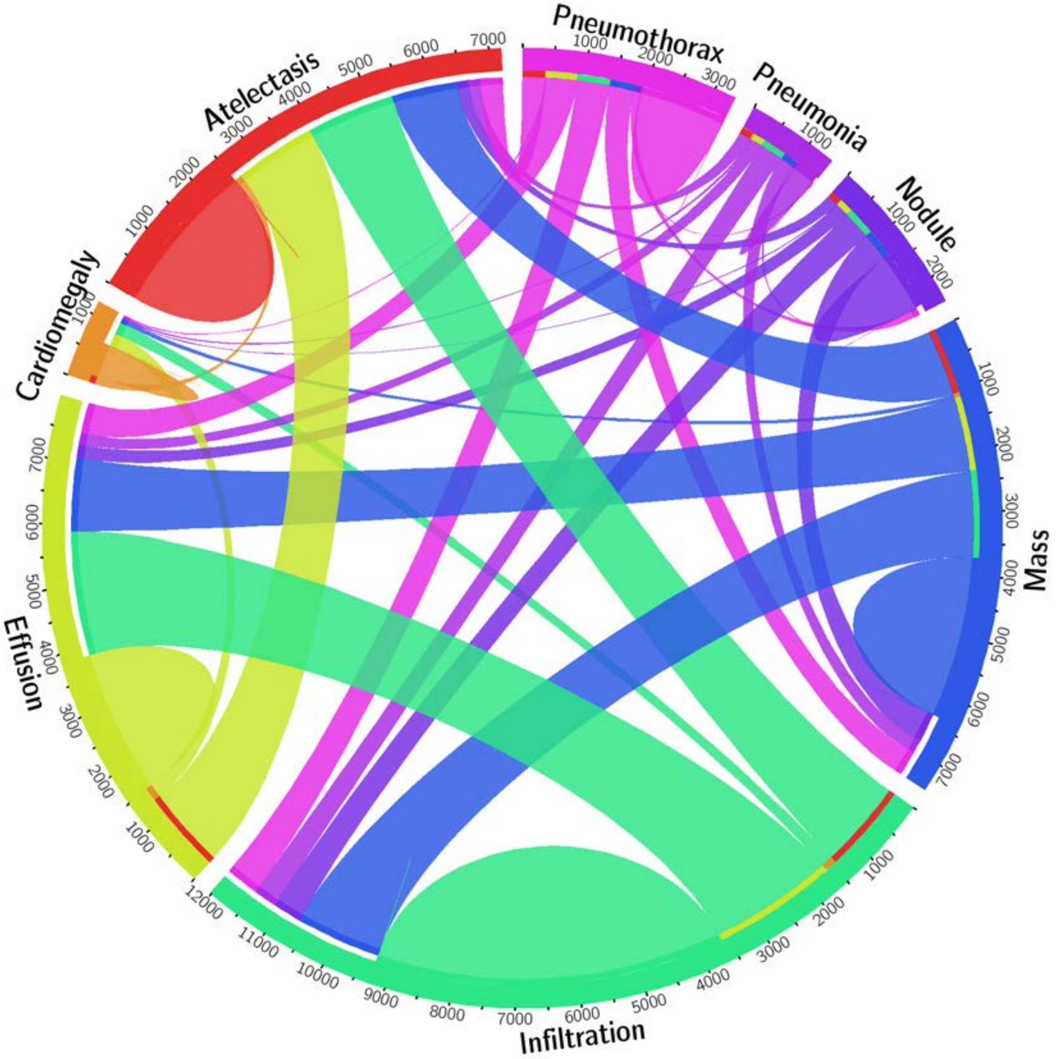


Image preparation

1. Select frontal view images only
2. Convert to 8-bit RBG images by using default window level and width
3. Resize image to 1024 * 1024
4. Totally 108,948 frontal view chest X-ray images of 32,717 unique patients. This is about 27 times of the previously available largest frontal view chest x-ray database OPENi.

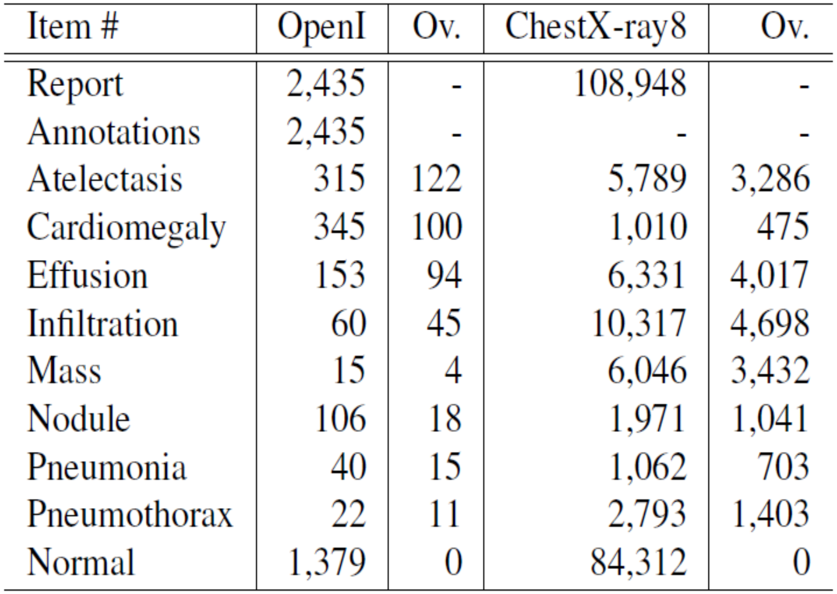
Disease Category Statistics

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

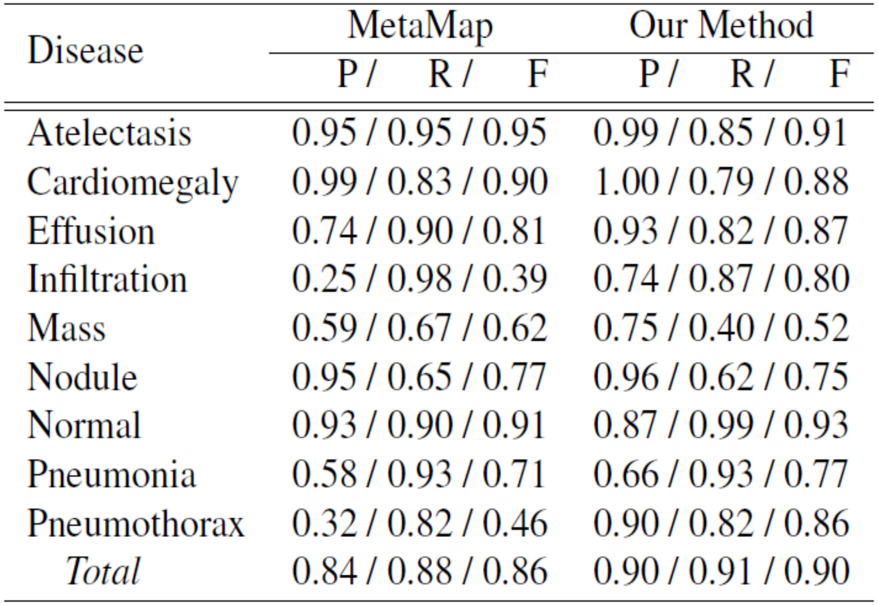


Evaluation of Disease Labeling

OPEN i®



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



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

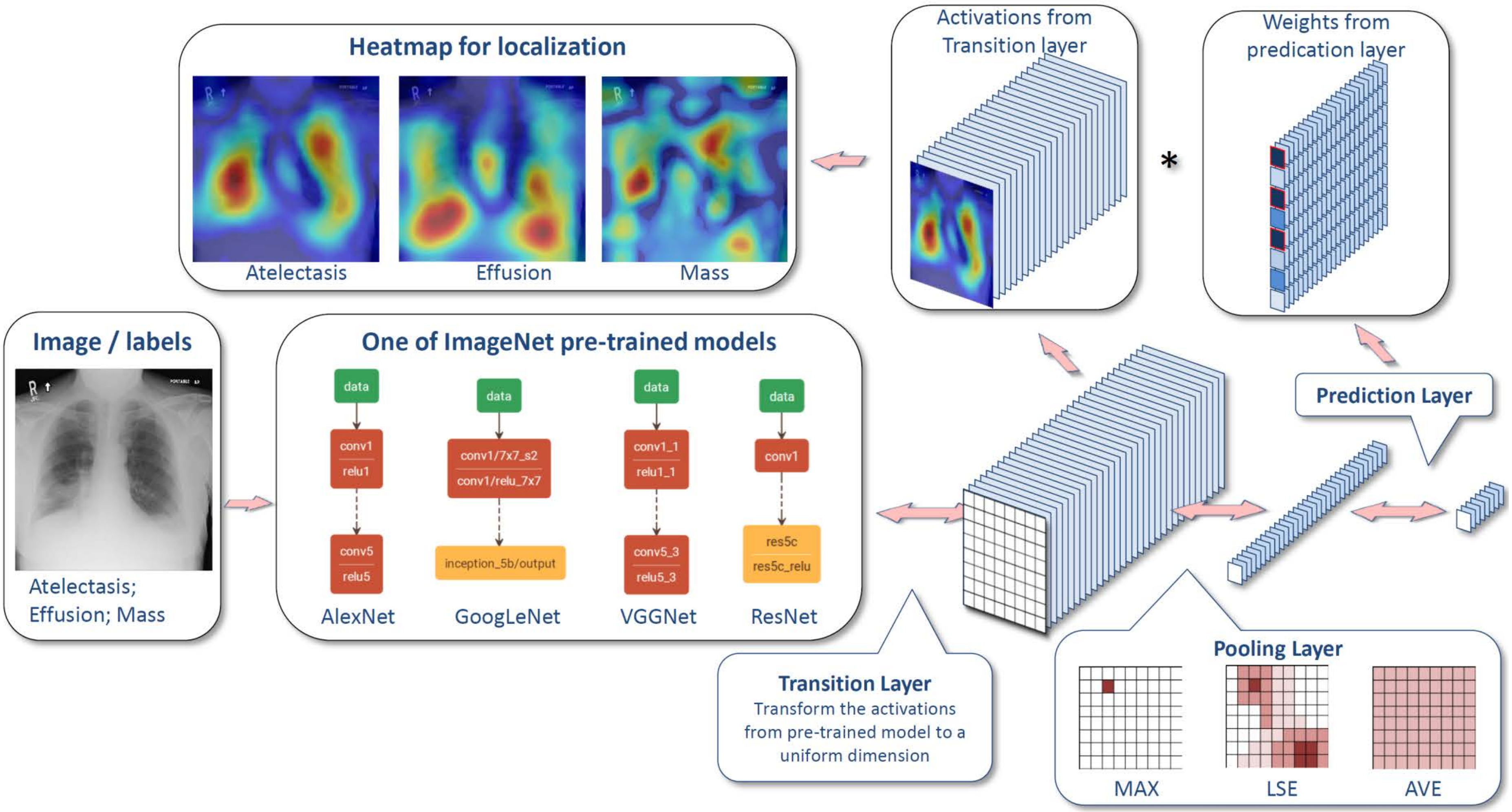
Table 2. *Evaluation of image labeling results on OpenI dataset. Performance is reported using P, R, F1-score.*

Weakly-Supervised Classification and Localization of Common Thorax Diseases

CVPR
July 21-26
HONOLULU 2017

07/21/2017

Framework Overview



Multi-label Setting

- We define a 8-dimensional label vector

$$\mathbf{y} = [y_1, ..., y_c, ..., y_C], y_c \in \{0, 1\}, C = 8$$

- y_c indicates the presence with respect to according pathology in the image
- A all-zero vector represents the status of “Normal”.
- This definition transits the multi-label classification problem into a regression-like loss setting

Transition Layer

- A variety of CNN networks are adopted and integrated into the proposed framework, e.g. GoogLeNet and ResNet
- Transform the activations from previous layers into a uniform dimension of output $S \times S \times D, S \in \{8, 16, 32\}$
- Pass down the weights from pre-trained DCNN models in a standard form, which is critical for using this layers' activations to further generate the heatmap in pathology localization step

Multi-label Classification Loss Layer

- We first experiment 3 standard loss functions for the regression task instead of using the softmax loss for traditional multi-class classification model, i.e., Hinge Loss (HL), Euclidean Loss (EL) and Cross Entropy Loss (CEL)
- The model has difficulty learning positive because of one-hot-like image labeling strategy (sparse image labels) and the unbalanced numbers of pathology and “Normal” classes.
- Positive/negative balancing, e.g.

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

Global Pooling Layer

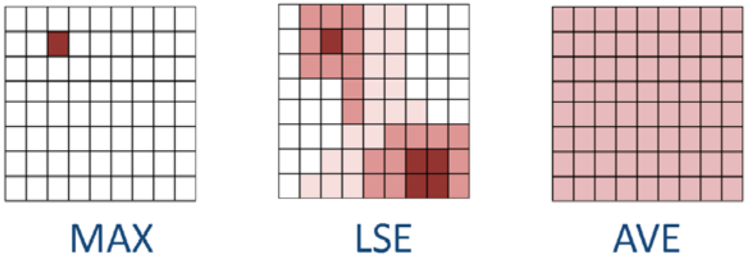
- The pooling layer plays an important role that chooses what information to be passed down [Zhou et al., 2016].
- Besides the conventional max pooling and average pooling, we also utilize the Log-Sum-Exp (LSE) pooling, which is defined as

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

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

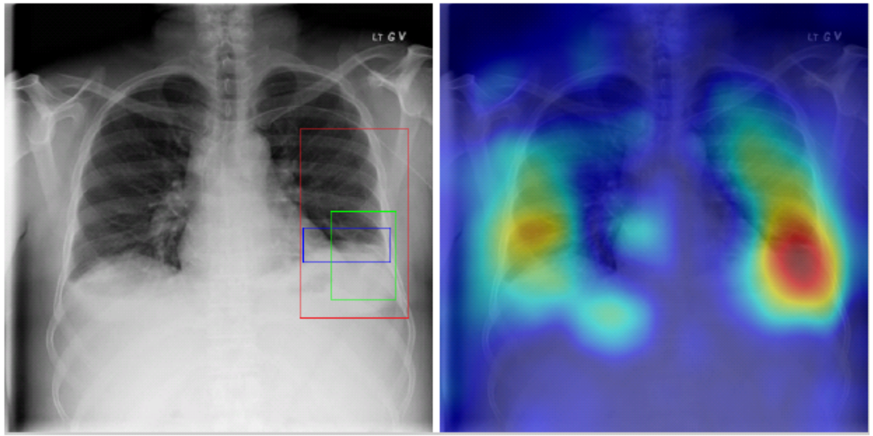
where $x^* = max\{|x_{ij}|, (i,j) \in \mathbf{S}\}$.

- By controlling the hyper-parameter r, the pooled value ranges from the maximum in S (when r -> 1) to average (r -> 0).



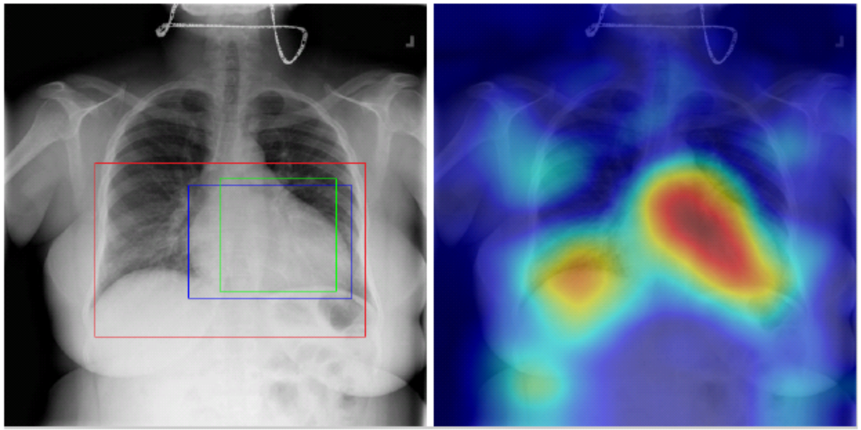
Disease Localization Results - Atelectasis

Radiology report	Keyword	Localization Result

findings include: 1. left basilar atelectasis/consolidation. 2. prominent hilum (mediastinal adenopathy). 3. left pic catheter (tip in atriocaval junction). 4. stable, normal appearing cardiomediastinal silhouette.
impression: small right pleural effusion otherwise stable abnormal study including left basilar infiltrate/atelectasis, prominent hilum, and position of left pic catheter (tip atriocaval junction).
	Effusion;
	Infiltration;
	Atelectasis
		

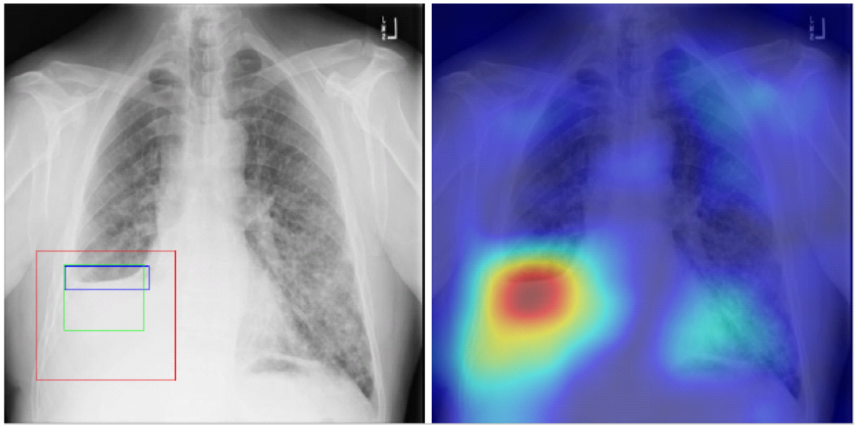
*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Cardiomegaly

Radiology report	Keyword	Localization Result
findings include: 1. cardiomegaly (ct ratio of 17/30). 2. otherwise normal lungs and mediastinal contours. 3. no evidence of focal bone lesion. dictating	Cardiomegaly	

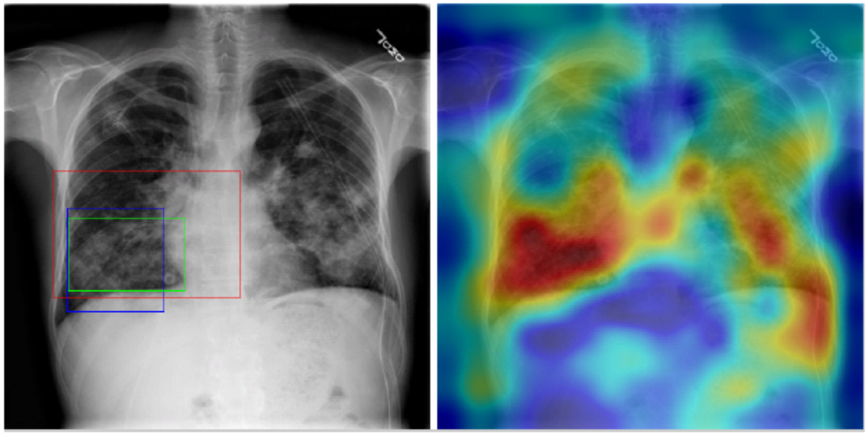
*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Effusion

Radiology report	Keyword	Localization Result
findings: no appreciable change since XX/XX/XX. small right pleural effusion. elevation right hemidiaphragm. diffuse small nodules throughout the lungs, most numerous in the left mid and lower lung. impression: no change with bilateral small lung metastases.	Effusion; Nodule	

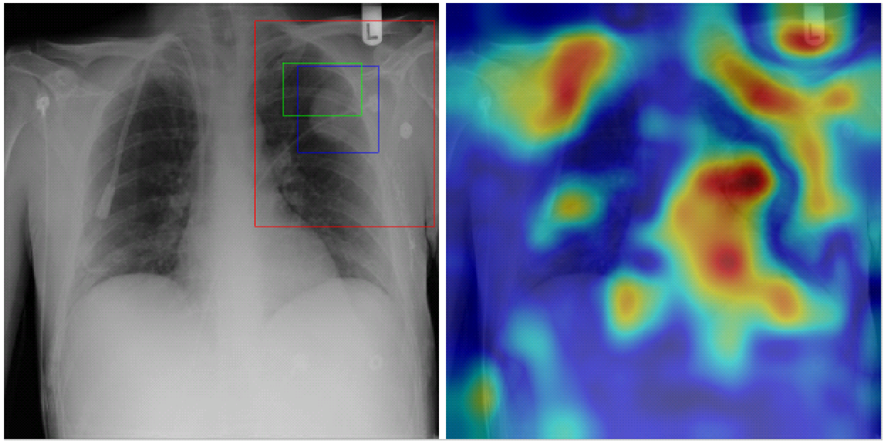
*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Infiltration

Radiology report	Keyword	Localization Result
findings: port-a-cath reservoir remains in place on the right. chest tube remains in place, tip in the left apex. no pneumothorax. diffuse patchy infiltrates bilaterally are decreasing. impression: infiltrates and effusions decreasing.	Infiltration	

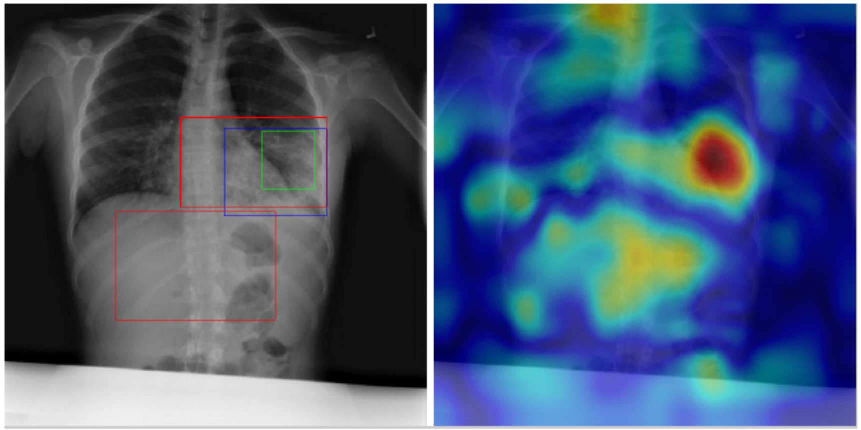
*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Mass

Radiology report	Keyword	Localization Result
findings: right internal jugular catheter remains in place. large metastatic lung mass in the lateral left upper lobe is again noted. no infiltrate or effusion. extensive surgical clips again noted left axilla. impression: no significant change.	Mass	

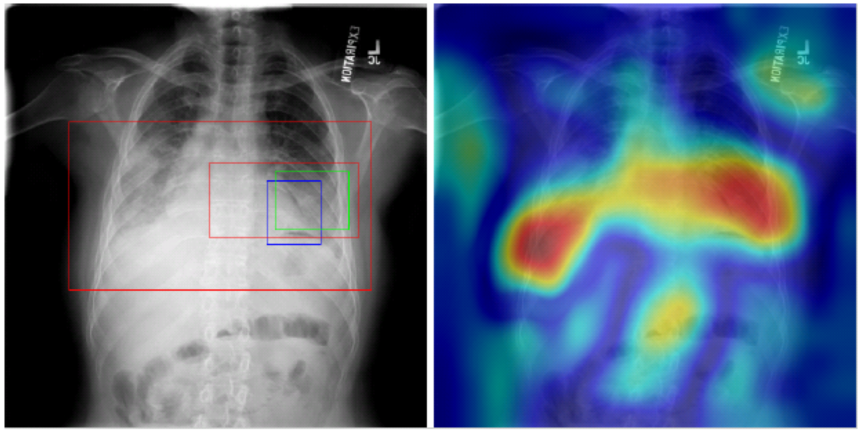
*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Pneumonia

Radiology report	Keyword	Localization Result
findings: unchanged left lower lung field infiltrate/air bronchograms. unchanged right perihilar infiltrate with obscuration of the right heart border. no evidence of new infiltrate. no evidence of pneumothorax the cardiac and mediastinal contours are stable. impression: 1. no evidence pneumothorax. 2. unchanged left lower lobe and left lingular consolidation/bronchiectasis. 3. unchanged right middle lobe infiltrate	Pneumonia; Infiltration	

*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Disease Localization Results - Pneumothorax

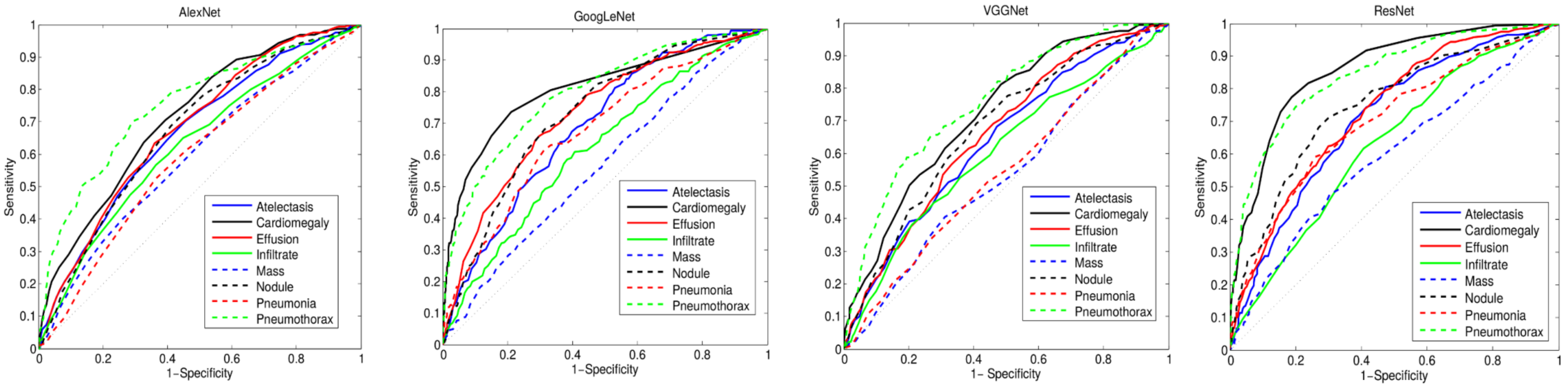
Radiology report	Keyword	Localization Result
findings: frontal lateral chest x-ray performed in expiration. left apical pneumothorax visible. small pneumothorax visible along the left heart border and left hemidiaphragm. pleural thickening, mass right chest. the mediastinum cannot be evaluated in the expiration. bony structures intact. impression: left post biopsy pneumothorax.	Mass; Pneumothorax	

*Correct bounding box (in green), false positives (in red) and the ground truth (in blue)

Experiment Setting

- 108,948 frontal-view X-ray images, 24,636 images contain one or more pathologies + 10,000 images of “Normal”
- Randomly shuffled the dataset into three subgroups: i.e. training (70%), validation (10%) and testing (20%)
- Multi-label CNN architecture is implemented using Caffe framework
- The ImageNet pre-trained models, i.e., AlexNet, GoogLeNet, VGGNet-16 and ResNet-50 are obtained from the Caffe model zoo
- Due to the large image size and the limit of GPU memory, reduce the image batch size while increasing the iter size to accumulate the gradients. We set batch size * iter size = 80

Multi-label Disease Classification Results

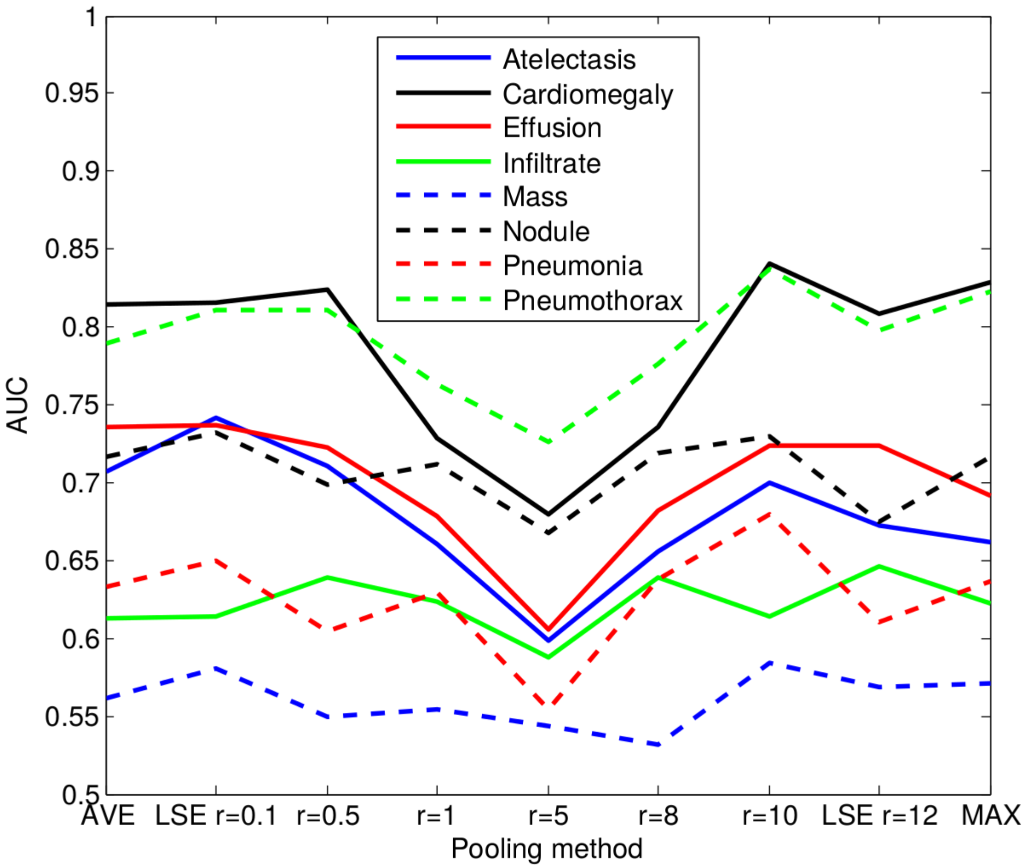


Setting	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
Initialization with different pre-trained models								
AlexNet	0.6458	0.6925	0.6642	0.6041	**0.5644**	0.6487	0.5493	0.7425
GoogLeNet	0.6307	0.7056	0.6876	0.6088	0.5363	0.5579	0.5990	0.7824
VGGNet-16	0.6281	0.7084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

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

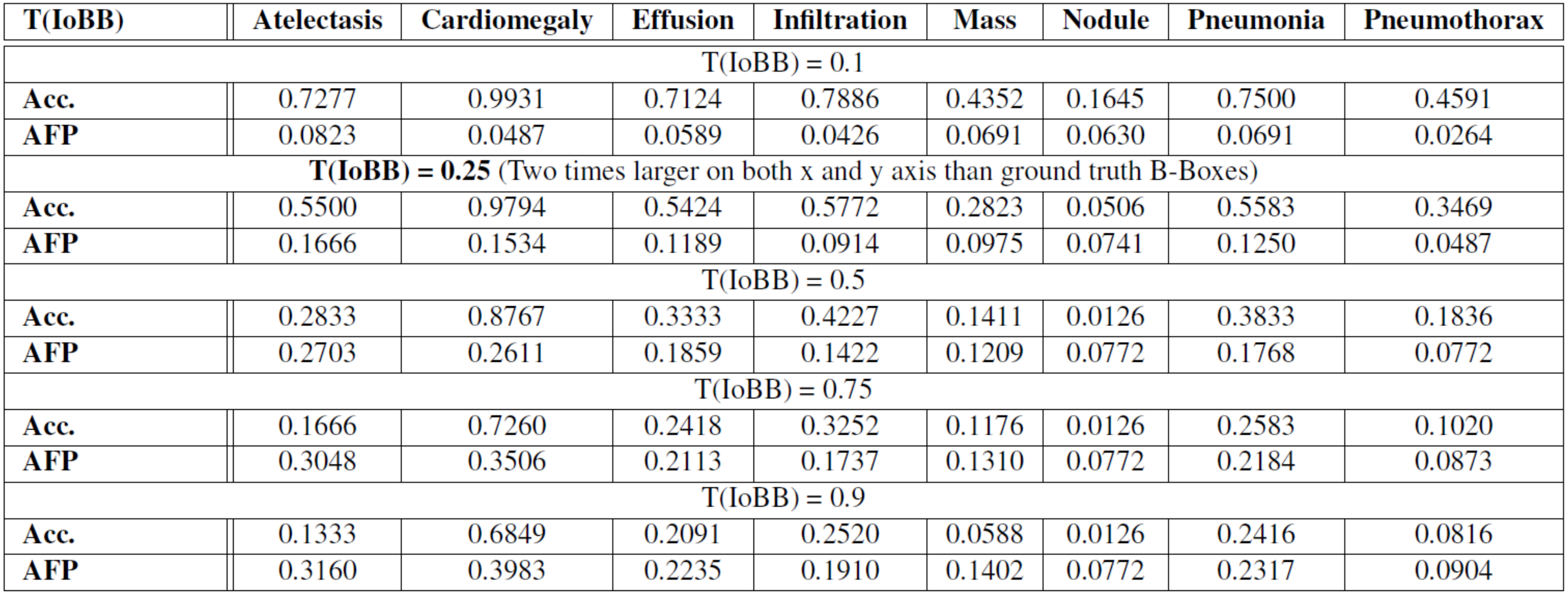
Note that the above Multi-label disease classification results have been noticeably improved since the CVPR deadline in Nov. 2016.

Different Spatial Pooling Strategies



- The hyper-parameter r in LSE pooling varies in { 0.1, 0.5, 1, 5, 8, 10, 12}
- LSE pooling behaves like a weighed pooling method or a transition scheme between average and max pooling under different r values

Disease Localization Results – IoBB*

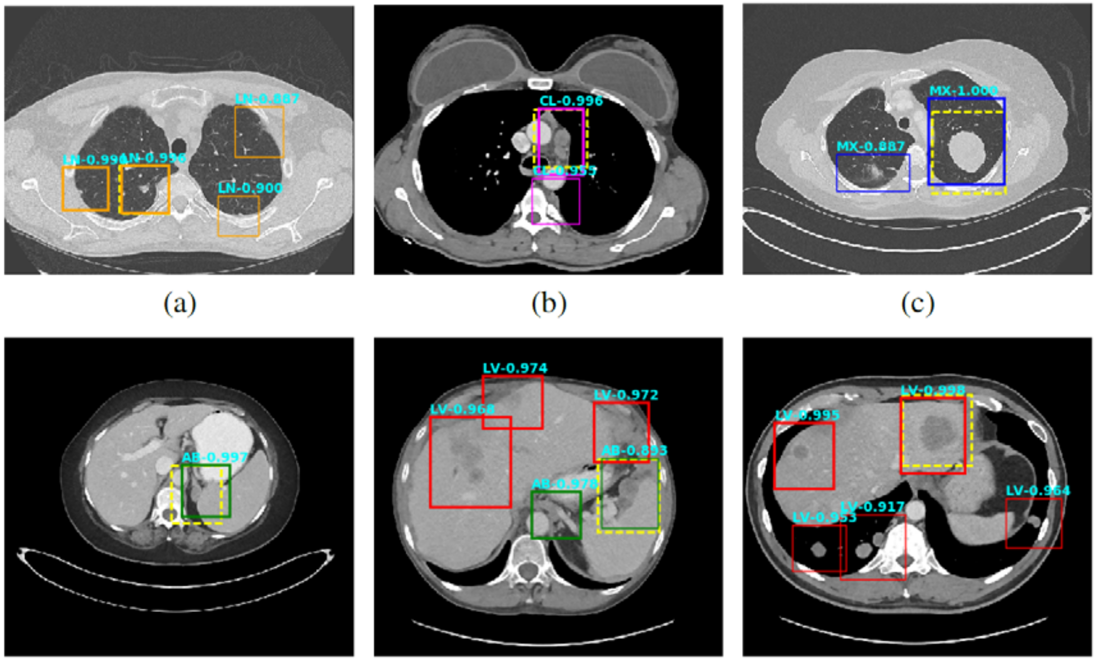


T(IoBB)	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
T(IoBB) = 0.1								
Acc.	0.7277	0.9931	0.7124	0.7886	0.4352	0.1645	0.7500	0.4591
AFP	0.0823	0.0487	0.0589	0.0426	0.0691	0.0630	0.0691	0.0264
T(IoBB) = 0.25 (Two times larger on both x and y axis than ground truth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

*Intersection over the detected B-Box area ratio (IoBB) (similar to Area of Precision or Purity)

Future work - Challenges

- Improving image labeling accuracy
  - Disease category
  - NLP technique
  - Annotate reports for evaluation
- Improving multi-label classification accuracy
  - Other CNN architecture, e.g. classic layer setting
  - More accurate image labels lead to more effective learning
  - LSTM text-assisted end-to-end deep training
- Improving localization accuracy
  - Better bounding box generation method
  - integrate text info. by adopting attention model & location information



(a) (b) (c)

(d) (e) (f)

Fig. 4. Six sample detection results are illustrated with the annotation lesion patches as yellow dashed boxes. The outputs from our proposed detection framework are shown in colored boxes with LiVer lesion (LV) in Red, Lung Nodule (LN) in Orange, ABdomen lesion (AB) in Green, Chest Lymph node (CL) in magenta and other MiXed lesions (MX) in blue. (a) Four lung lesions are all correctly detected; (b) Two lymph nodes in mediastinum is presented; (c) A Ground Glass Opacity (GGO) and a mass are detected in the lung; (d) An adrenal nodule; (e) Correct detections on both the small abdomen lymph node nearly aorta but also other metastases in liver and spleen; (f) Two liver metastasis are correctly detected. Three lung metastases are detected but erroneously classified as liver lesions .

Runtime of 88ms to label a testing 512x512 slice on a single Titan-X GPU, → ~1 minute to read a 700 slice Chest/Abdomen CT scan!

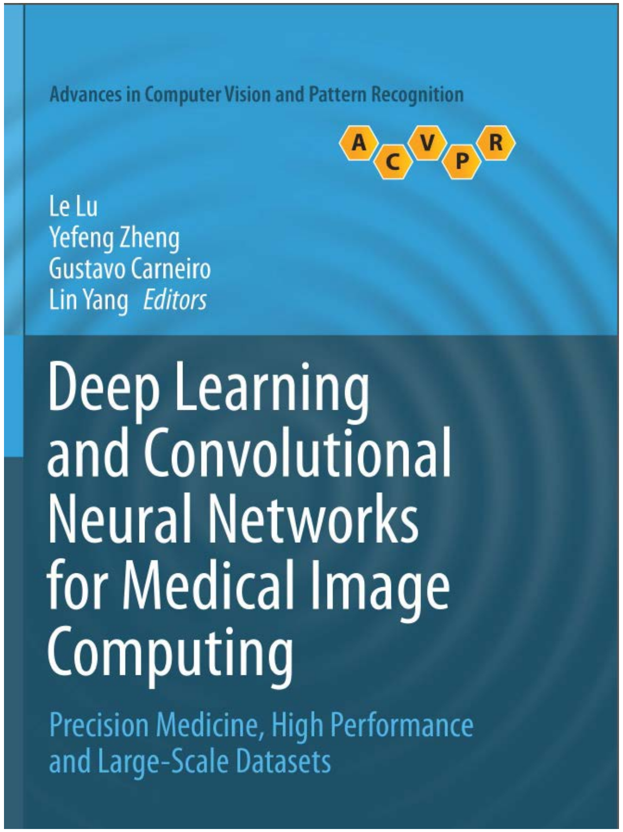
A new CAD paradigm via mining ultra-large scale retrospective clinical datasets with weak annotations: → universal, multi-purpose deeply trainable CAD systems, almost effortlessly from the workload perspective required for radiologists or human annotators.

(Patent Pending)

Database #3?

The day of "big data, weak label, true clinical impacts" seems coming faster than what we all have realized, just a few months ago. Thanks to Deep Learning developments in Computer Vision, MICCAI and "Open Science Initiative!"

So get prepared! Exciting days are yet to come …



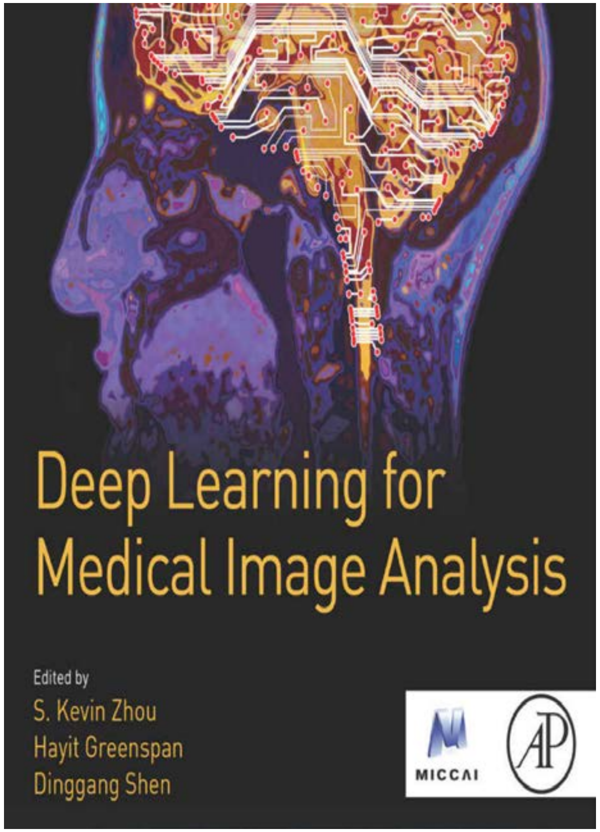
Advances in Computer Vision and Pattern Recognition

A C V P R

Le Lu
Yefeng Zheng
Gustavo Carneiro
Lin Yang *Editors*

Deep Learning and Convolutional Neural Networks for Medical Image Computing

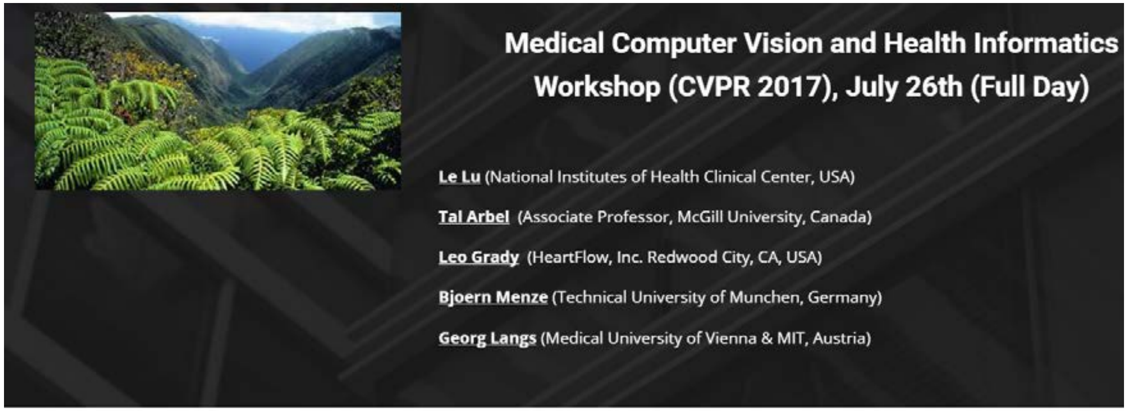
Precision Medicine, High Performance and Large-Scale Datasets



Deep Learning for Medical Image Analysis

Edited by
S. Kevin Zhou
Hayit Greenspan
Dinggang Shen

You are cordially invited to be part of the force!



Medical Computer Vision and Health Informatics Workshop (CVPR 2017), July 26th (Full Day)

Le Lu (National Institutes of Health Clinical Center, USA)

Tal Arbel (Associate Professor, McGill University, Canada)

Leo Grady (HeartFlow, Inc. Redwood City, CA, USA)

Bjoern Menze (Technical University of Munchen, Germany)

Georg Langs (Medical University of Vienna & MIT, Austria)

Confirmed Speakers (and growing ..., We are preparing up to 16 high quality talks of 30 minutes each):

1, Prof. Daniel Rueckert (Imperial College London)

2, Prof. Jiebo Luo (University of Rochester)

3, Prof. Chunhua Shen (University of Adelaide)

4, Prof. Greg Hager (Johns Hopkins University)

5, Prof. Pascal Fua (EPFL)

6, Prof. Nikos Paragios (Ecole Centrale Paris)

7, Dr. Dale Webster, Dr. Martin Stumpe (Google Research)

8, Dr. Xiaosong Wang (NIH Clinical Center)

9, Prof. Junzhou Huang (UT Arlington)

10, Prof. Lin Yang (University of Florida)

11, Prof. Zhaozheng Yin (Missouri University S&T)

12, Mr. Andre Esteva (Stanford University)

**Bridging CVPR, ICCV, MICCAI, ECCV, NIPS, ISBI, RSNA, GTC … to solve important medical (imaging) diagnosis problems for the life-saving mission!

CVPR July 21-26 HONOLULU 2017

THANK YOU!

Scan to download the Key Radiology Image and Chest X-ray datasets via Google Cloud (free to everyone!)

https://console.cloud.google.com/storage/gcs-public-data--nih

Acknowledgement

- This work is supported by the Intramural Research Program of the National Institutes of Health Clinical Center and National Library of Medicine.
- We thank Nvidia corporation for the GPU donation.

NIH National Institutes of Health
Turning Discovery Into Health

Scan to contact

Thank you & our amazing trainees, collaborators, Industrial partnerships!

Department of Radiology and Imaging Sciences
National Institutes of Health Clinical Center
Bethesda, Maryland 20892-1182

Contacts: le.lu@nih.gov; rms@nih.gov

Thanks go to NIH Intramural Research Program (NIH-IRP) for great support, NVidia for donating Tesla K40 and Titan X GPUs! Six NIH FARE awards (2014,2015, 2016), KRIBB/NSERC/DNSEG/ISTP Fellowships, NIH-PingAn CRADA Collaboration Program, MICCAI student travel award 2016, RSNA trainee research prize 2016, …