Big Data, Weak Label and True Clinical Impacts for Radiology Imaging Diagnosis

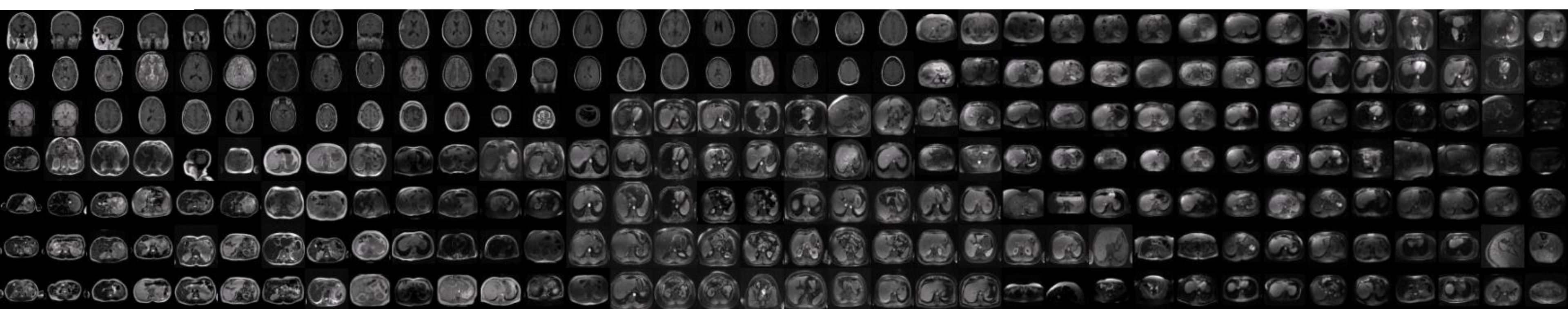
Xiaosong Wang

xiaosong.wang@nih.gov

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

CVPR July 21-26 2017 HONOLULU

Medical Computer Vision and Health Informatics Workshop



NIH National Institutes of Health
Turning Discovery Into Health

Motivation I

- The availability of well-labeled data is the key for large scale machine learning, e.g. deep learning
- Different levels of annotations are required for a variety of medical imaging problems.

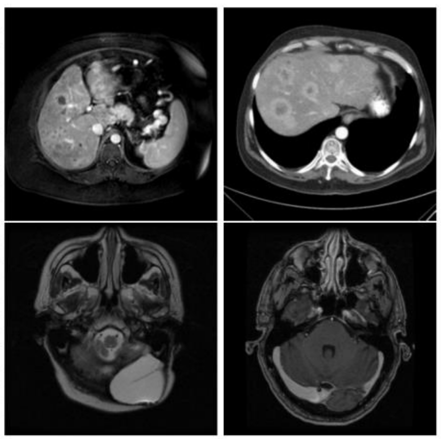
Classification (image label)

Cluster #23

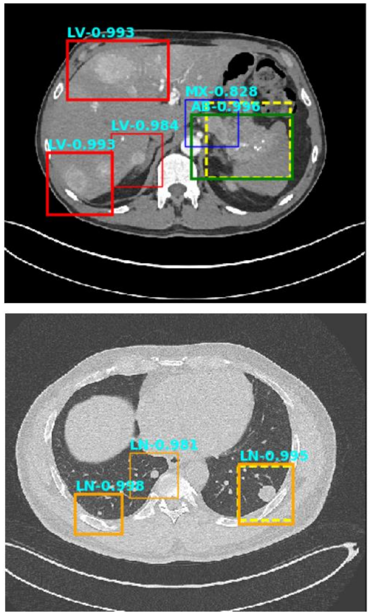
Word	Frequency
liver	524
abdomen	337
enhancement	217
mass	198
lesion	168
lobe	161
adenopathy	119
lesions	109
segment	58
bulky	45

Cluster #64

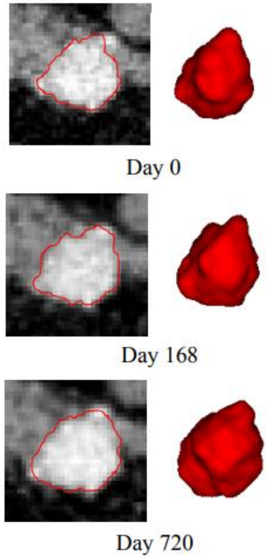
Word	Frequency
enhancement	277
cerebellar	193
lesion	192
lobe	186
flair	173
hemisphere	155
mass	134
abnormal	119
frontal	115
cerebellum	113



Localization(b-box)



Segmentation (mask) *



Ground truth of tumor growth at different time points.

* Image Credit: Ling Zhang et. al, "Personalized Pancreatic Tumor Growth Prediction via Group Learning", MICCAI 2017

National Institutes of Health
Turning Discovery Into Health

Motivation II

- High quality labels for large medical imaging database are NOT available
- Annotation on medical images usually requires professionals with clinical training.
- Conventional ways for collecting image labels are NOT applicable, e.g.
  - ❑ Internet search followed by crowd-sourcing

Large scale natural image datasets

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

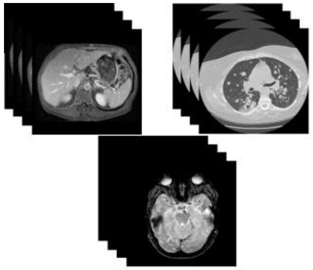
Large scale *<u>Medical Image</u>* *dataset*

?

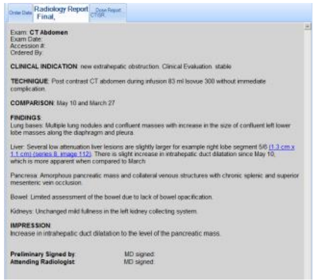
* Dataset logos shown here are from respective public dataset websites.

Where To Dig?

Hospital PACS



Images (different patient across time)



Radiology Report Final,

Exam: **CT Abdomen**
Exam Date:
Accession #:
Ordered By:

CLINICAL INDICATION: new extrahepatic obstruction. Clinical Evaluation: stable

TECHNIQUE: Post contrast CT abdomen during infusion 83 ml Isovue 300 without immediate complication.

COMPARISON: May 10 and March 27

FINDINGS:
Lung bases: Multiple lung nodules and confluent masses with increase in the size of confluent left lower lobe masses along the diaphragm and pleura.

Liver: Several low attenuation liver lesions are slightly larger for example right lobe segment 5/6 (1.3 cm x 1.1 cm) (series 8, image 112). There is slight increase in intrahepatic duct dilatation since May 10, which is more apparent when compared to March.

Pancreas: Amorphous pancreatic mass and collateral venous structures with chronic splenic and superior mesenteric vein occlusion.

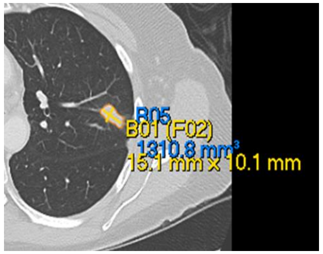
Bowel: Limited assessment of the bowel due to lack of bowel opacification.

Kidneys: Unchanged mild fullness in the left kidney collecting system.

IMPRESSION:
Increase in intrahepatic duct dilatation to the level of the pancreatic mass.

Preliminary Signed by: MD signed:
Attending Radiologist: MD signed:

Associated report (Findings,attribues)



Clinical Annotation (measure, location, lesion contour)

... ...

?

GT Label for Learning

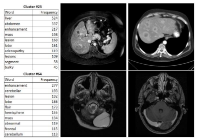
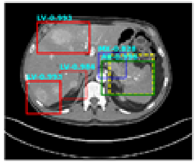
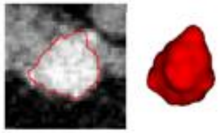


Image Label



Bounding box



Mask

... ...

Dig where doctors care!

Prospecting…

I. Unsupervised Joint Mining of Deep Features and Image Labels:
   - Hypothesised "Convergence": better labels lead to better trained Convolutional Neural Network (CNN) models which consequently feed more effective deep image features to facilitate more meaningful clustering/labels.
   - Clinical Application: image categorization / classification

II. Mining of Radiology Reports via NLP:
   - A two-stage process: pathology detection plus negation and uncertainty elimination.
   - Clinical Application: disease classification / localization

III. Utilizing Clinical Annotation as Weak Supervision:
   - Annotations suggest location information
   - Clinical Application: disease detection

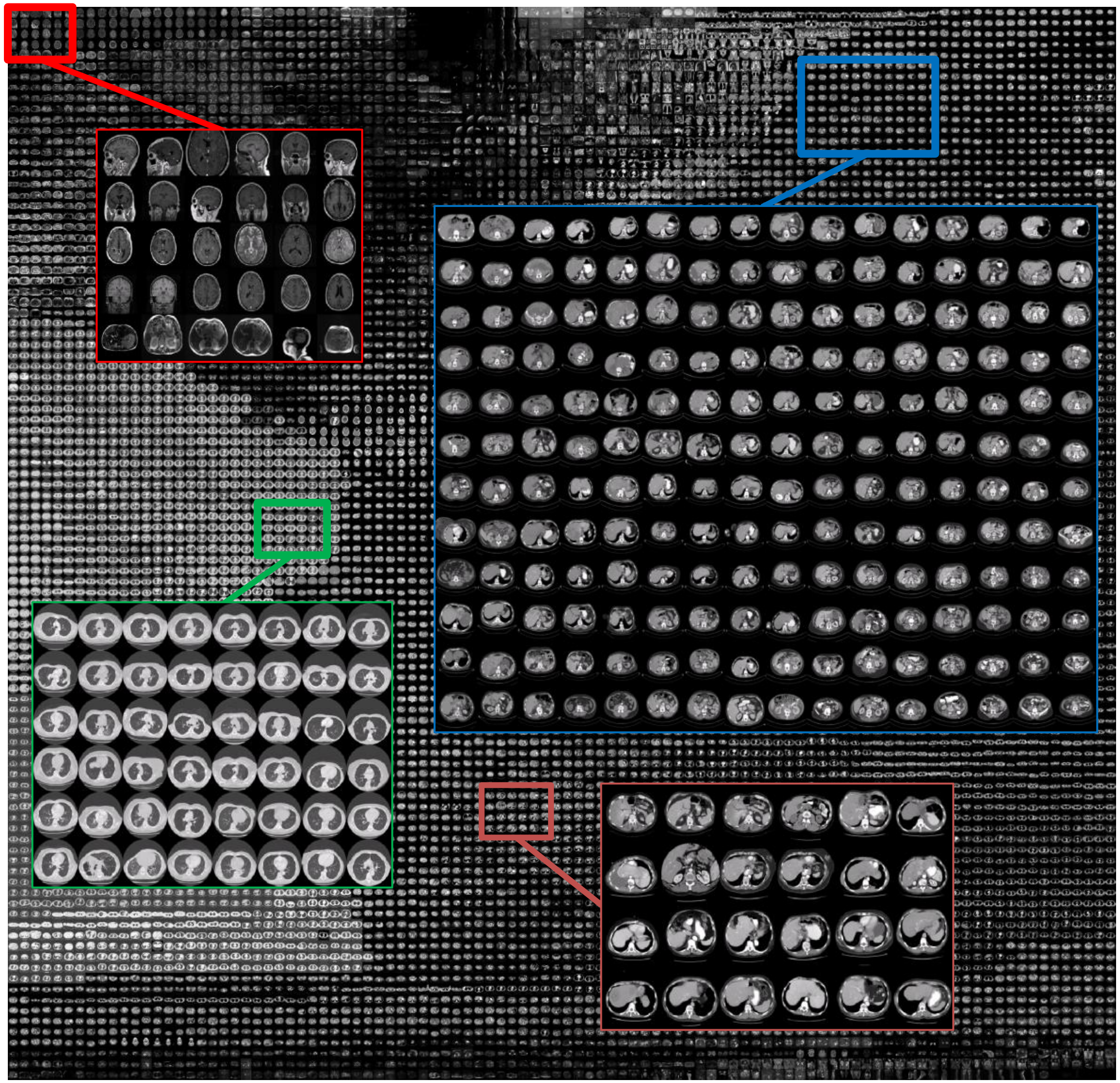
Case Study 1
Unsupervised Categorization of Images

X. Wang et al. Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Annotation and Scene Recognition. IEEE WACV, 2017

US Patent Application 62/302,096

Dataset

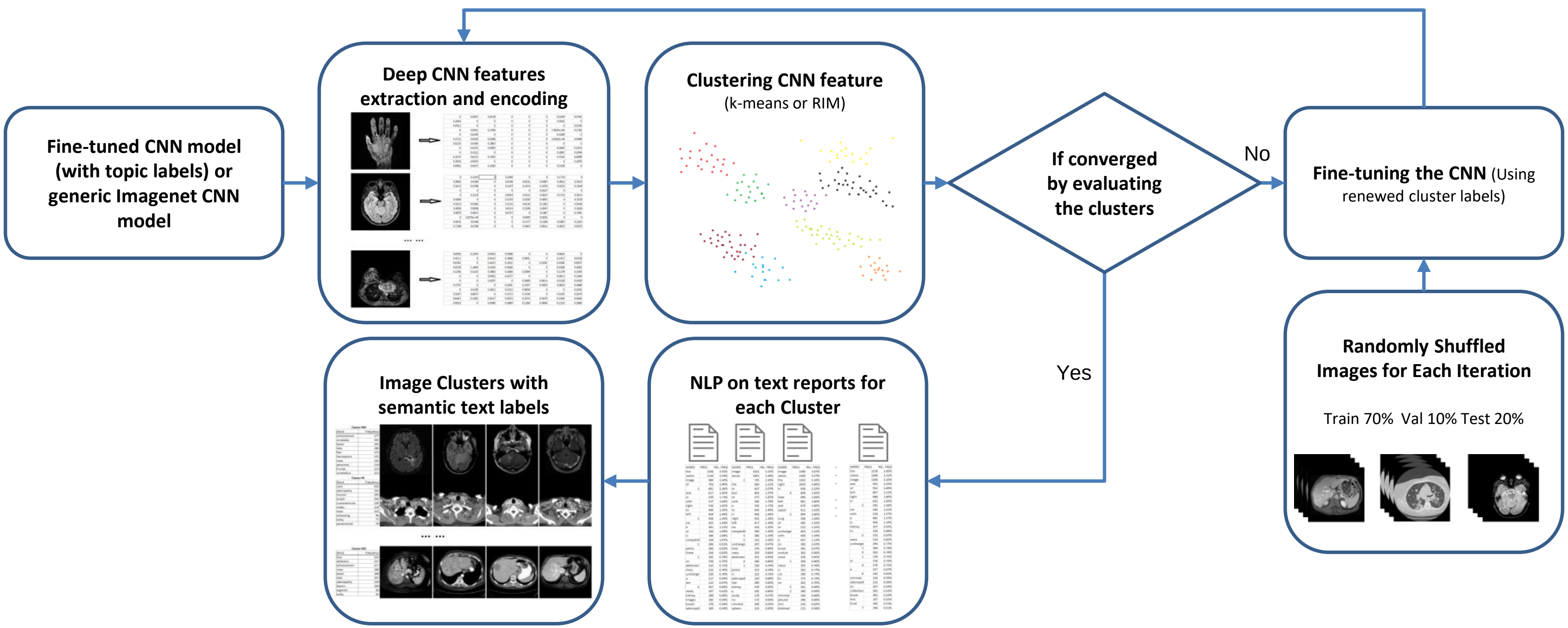
- “Keyimage” dataset: 215,786 key images from 17,845 unique patients.
- Key images are significant one or more images in a study referenced in the linked radiological report.
- Key images are directly extracted from the DICOM file and resized as 256*256 bitmap images (.png).
- Their intensity ranges 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

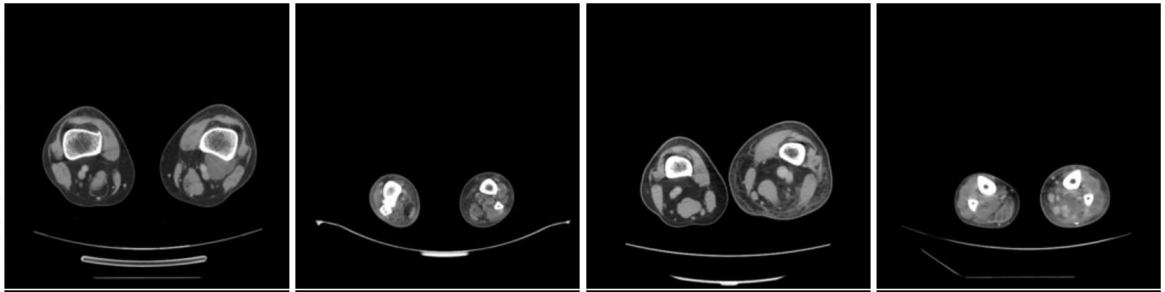


- Hypothesized “convergence”: better labels lead to better trained Convolutional Neural Network (CNN) models which consequently feed more effective deep image features to facilitate more meaningful clustering/labels.

Sample Categories

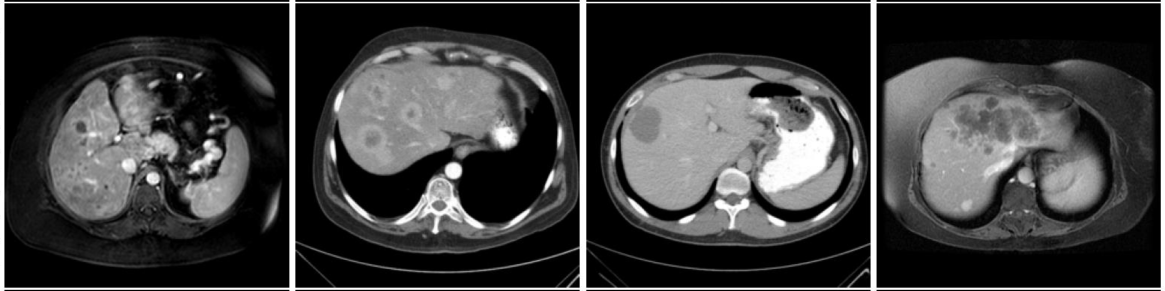
Cluster #14

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



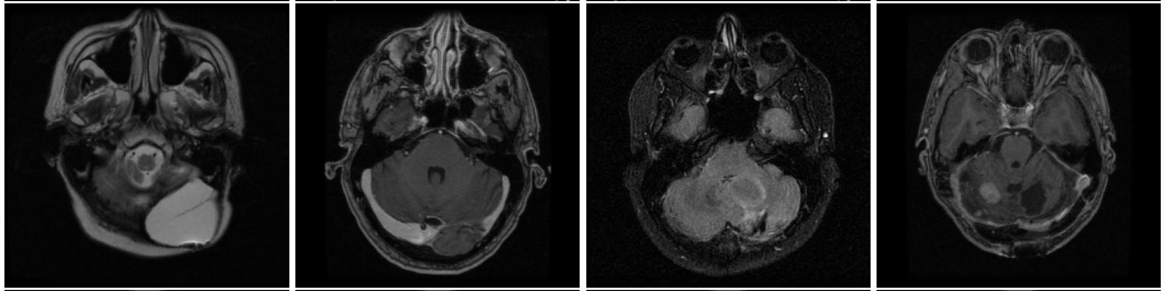
Cluster #23

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



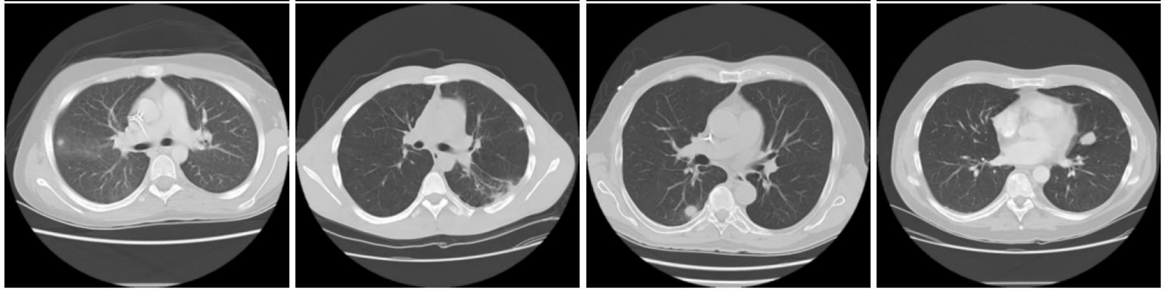
Cluster #64

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

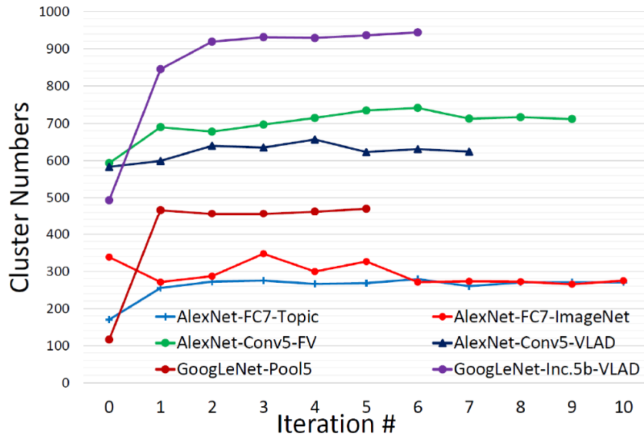


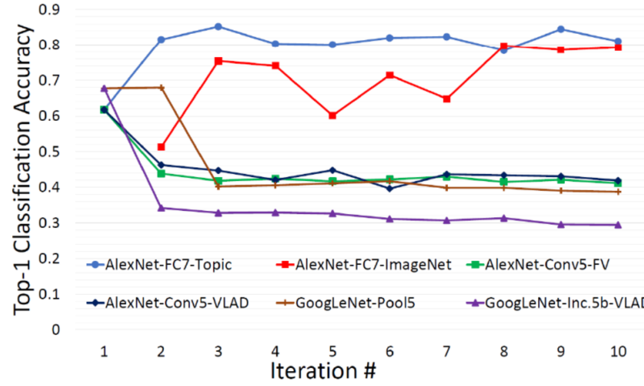
Experiment - Convergence

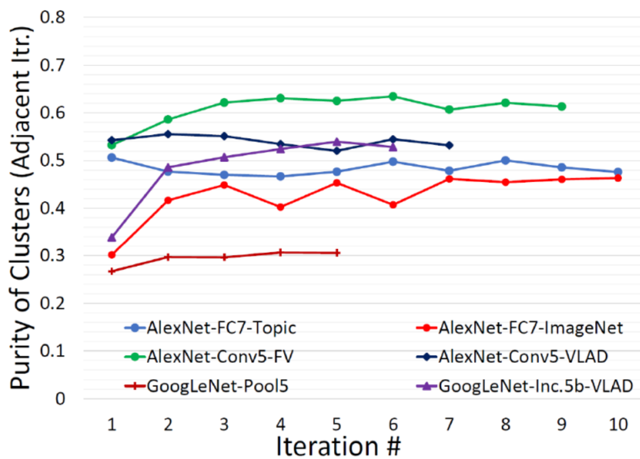
- Clustering via K-means only or over-fragmented K-means followed by Regularized Information Maximization (as an effective model selection method), are extensively explored and empirically evaluated.
- Two convergence measurements have been adopted, i.e., Clustering Purity and Normalized Mutual Information (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

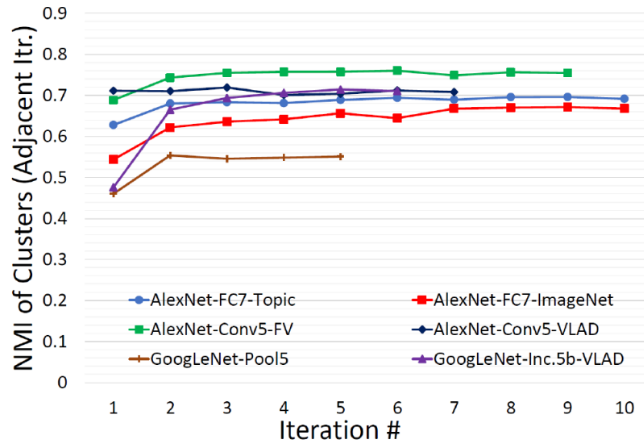
National Institutes of Health
Turning Discovery Into Health

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.

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

NIH National Institutes of Health
Turning Discovery Into Health

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

Case Study 2

Mining of Image Labels via NLP in Radiology Reports

X. Wang, Y. Peng, et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR, 2017

US Patent Application 62/476,029


NIH Chest X-ray

A Sample Radiology Report

Chest X-ray Radiology Report

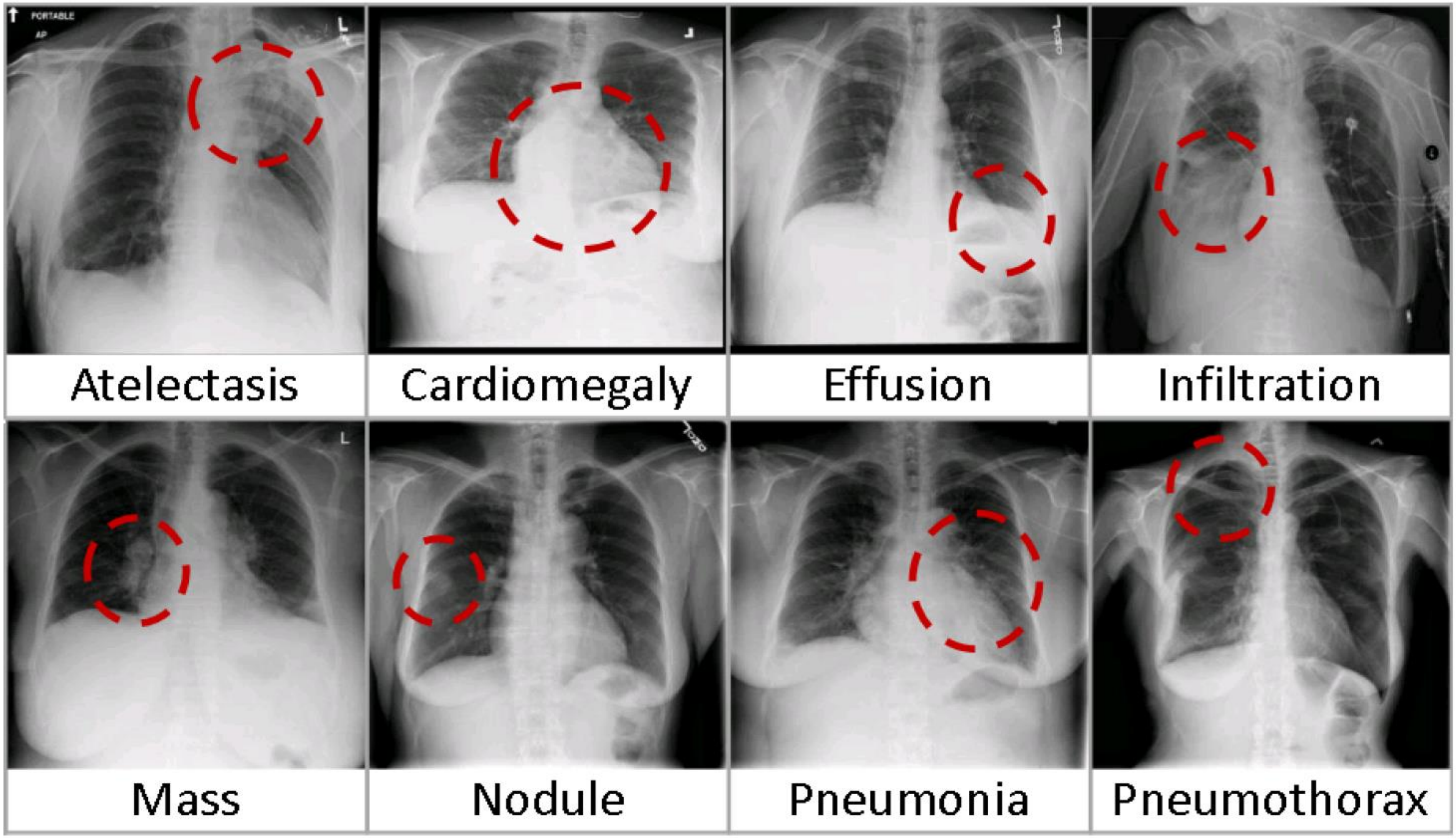
Findings:

unchanged <u>left lower lung </u>field **infiltrate**/air bronchograms. Unchanged <u>right perihilar</u> **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

8 Common Thorax Diseases



2-Stage Disease Label Mining

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

2-Stage Disease Label Mining

Stage 2: Removal of negation and uncertainty

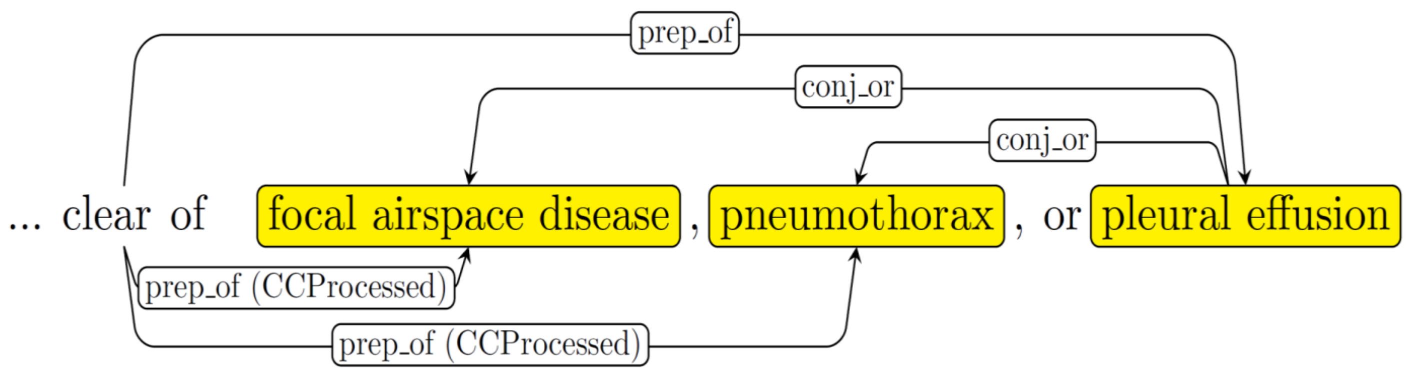
- Rule out those negated pathological statements and uncertain mentions of findings and diseases
- Defined the rules on the dependency graph, by utilizing the dependency label and direction information between words, e.g.

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

2-Stage Disease Label Mining

Stage 2: Removal of negation and uncertainty

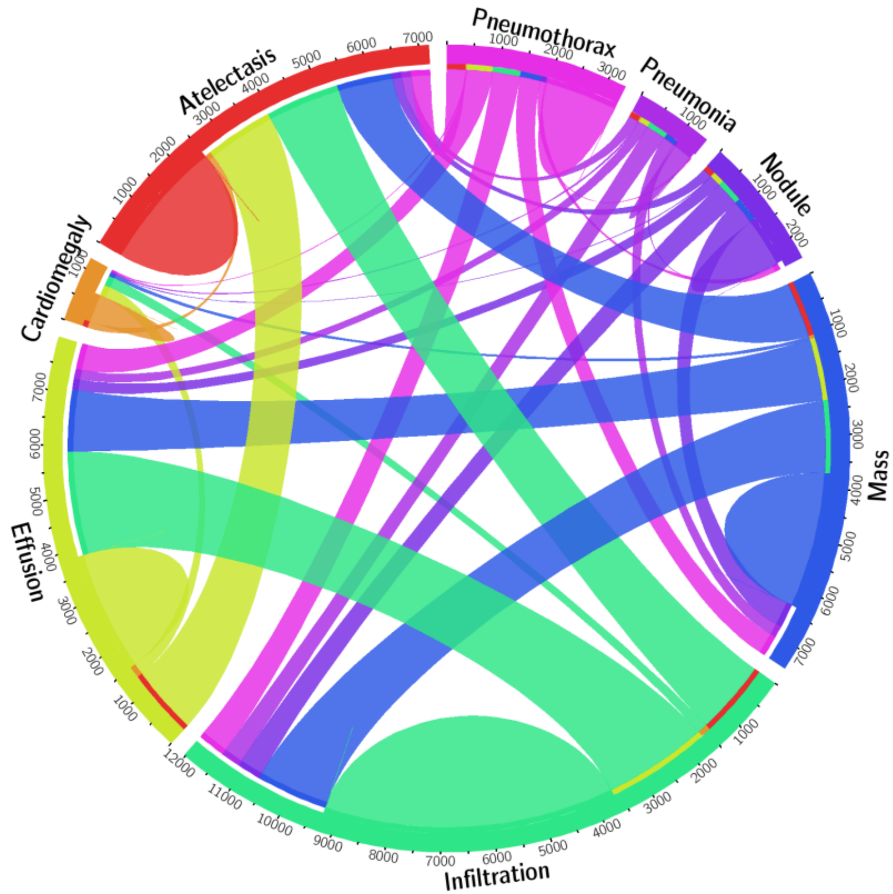
- Rule out those negated pathological statements and uncertain mentions of findings and diseases
- Defined the rules on the dependency graph, by utilizing the dependency label and direction information between words, e.g.



ChestX-ray Dataset 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

- # of images in each disease category with Overlay (Ov.) compared with OpenI



- Graphical illustration of correlations among diseases

Evaluation of Disease Labeling

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

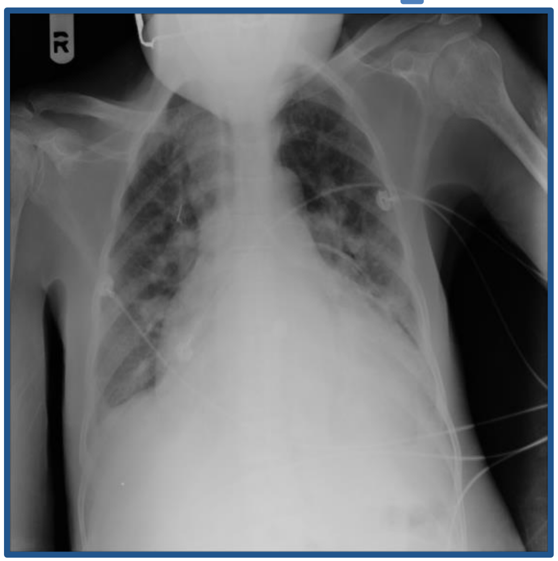
Table 2. *Evaluation of image labeling results on OpenI dataset. Performance is reported using P, R, F1-score.*

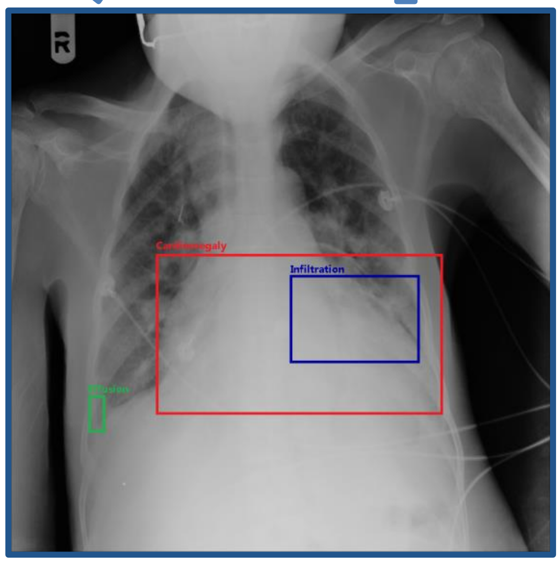
National Institutes of Health
Turning Discovery Into Health

Clinical Application

Multi-disease Classification and Localization

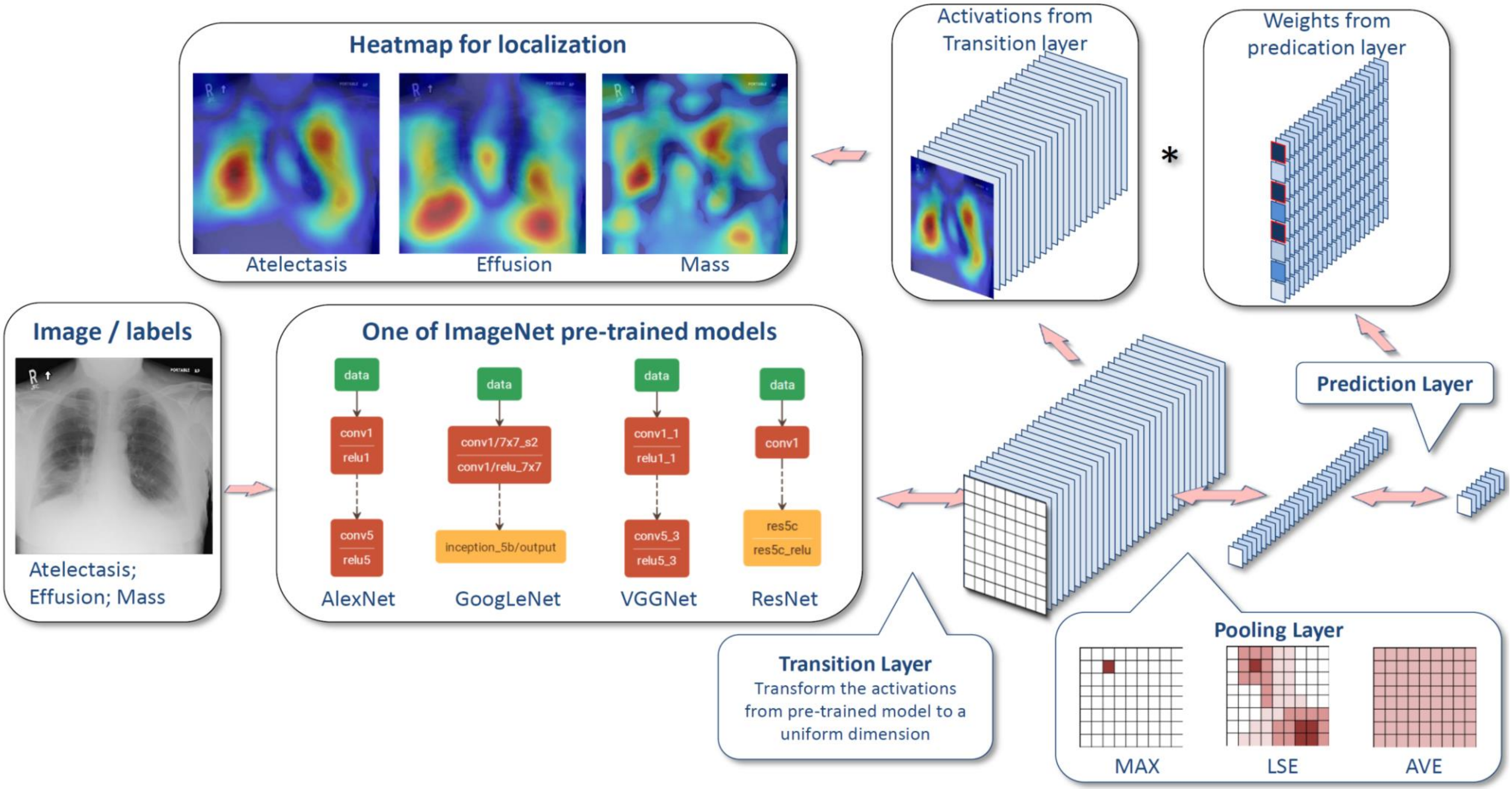
Finding Report Generation





FINDINGS: Heart and mediastinum remain markedly enlarged consistent with cardiomegaly. Basilar infiltrates and small pleural effusions.

Multi-label Classification and Localization Framework



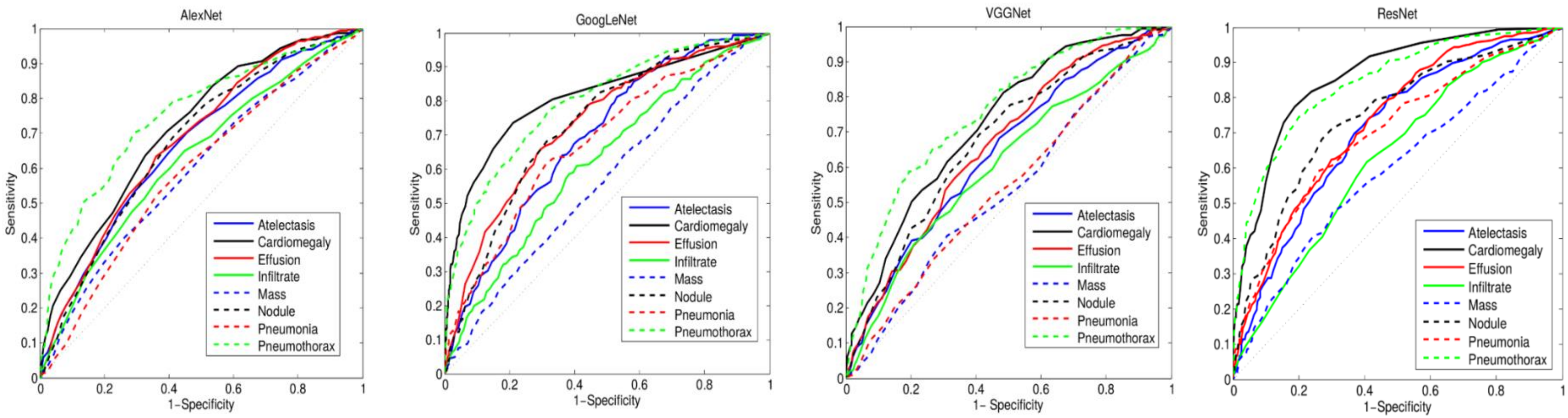
National Institutes of Health
Turning Discovery Into Health

Experiment Setting

- 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

National Institutes of Health
Turning Discovery Into Health

Multi-Disease Classification Results

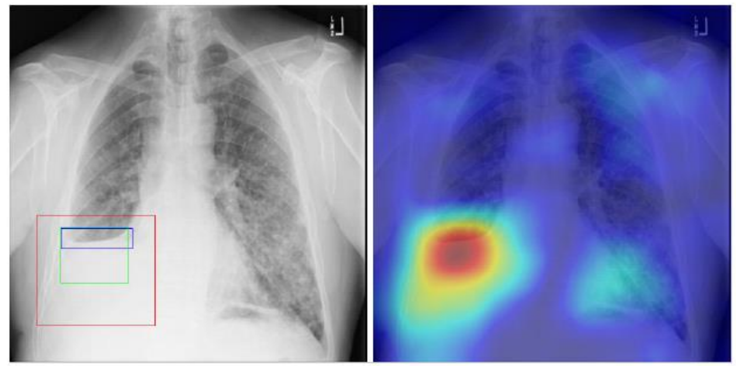


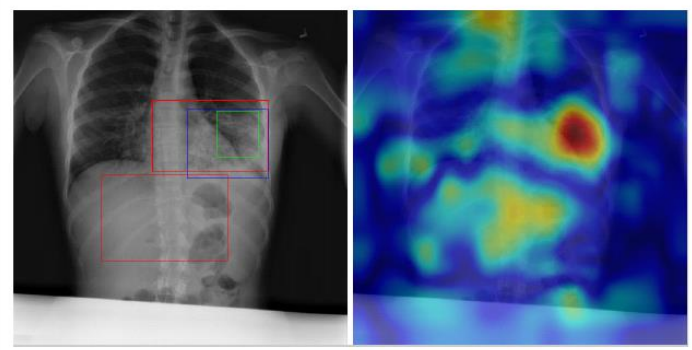
Setting	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
Initialization with different pre-trained models								
AlexNet	0.6458	0.6925	0.6642	0.6041	**0.5644**	0.6487	0.5493	0.7425
GoogLeNet	0.6307	0.7056	0.6876	0.6088	0.5363	0.5579	0.5990	0.7824
VGGNet-16	0.6281	0.7084	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.*

National Institutes of Health
Turning Discovery Into Health

Disease Localization Results

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	

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)

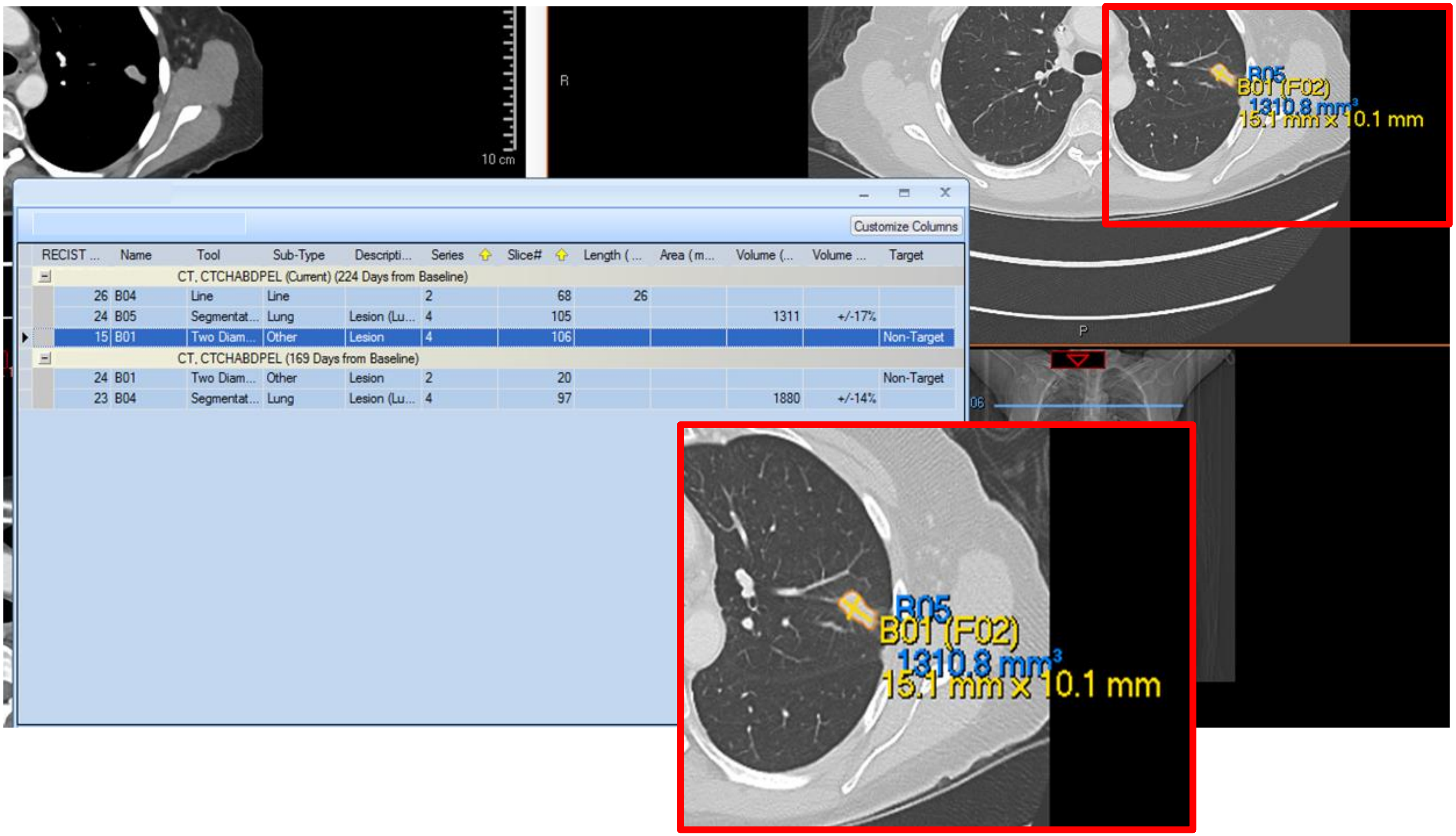
National Institutes of Health
Turning Discovery Into Health

Case Study 3
Utilizing Clinical Annotation as Weak Supervision

X. Wang, K. Yan et al. DeepLesion: Automated Deep Mining, Categorization and Detection of Significant Radiology Image Findings using Large-Scale Clinical Lesion Annotations, RSNA 2017

US Patent Application 62/514,223

Clinical Annotation Data



RECIST ...	Name	Tool	Sub-Type	Descripti...	Series	Slice#	Length (...	Area (m...	Volume (...	Volume ...	Target
CT, CTCHABDPEL (Current) (224 Days from Baseline)											
26	B04	Line	Line		2	68	26				
24	B05	Segmentat...	Lung	Lesion (Lu...	4	105			1311	+/-17%	
15	B01	Two Diam...	Other	Lesion	4	106					Non-Target
CT, CTCHABDPEL (169 Days from Baseline)											
24	B01	Two Diam...	Other	Lesion	2	20					Non-Target
23	B04	Segmentat...	Lung	Lesion (Lu...	4	97			1880	+/-14%	

Multi-Category Lesion Detection

- A new computer-aided detection paradigm with weak annotations mined from large scale retrospective clinical datasets.
- Different from traditional CADe system
- Multi-category multi-lesion detection in 3D volume
- Almost effortlessly from the workload perspective required for radiologists or human annotators.

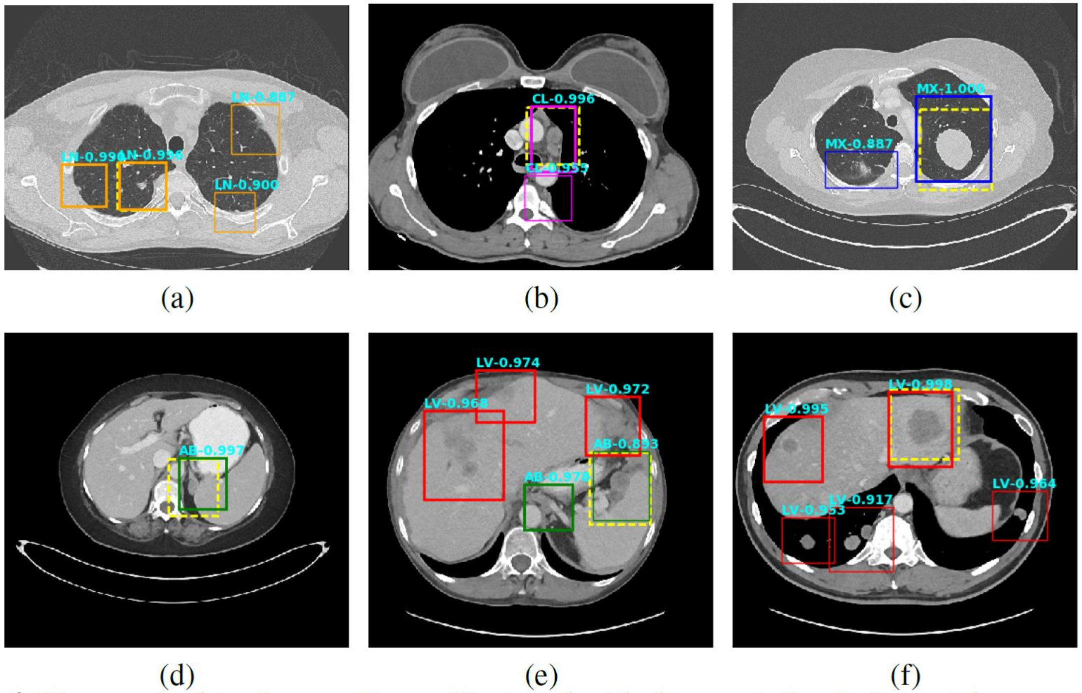


Fig. 4. Six sample detection results are illustrated with the annotation lesion patches as yellow dashed boxes. The outputs from our proposed detection framework are shown in colored boxes with LiVer lesion (LV) in red, Lung Nodule (LN) in orange, ABdomen lesion (AB) in green, Chest Lymph node (CL) in magenta and other MiXed lesions (MX) in blue. (a) Four lung lesions are all correctly detected; (b) Two lymph nodes in mediastinum are 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 near aorta but also other metastases in liver and spleen; (f) Two liver metastases are correctly detected. Three lung metastases are detected but erroneously classified as liver lesions.

National Institutes of Health
Turning Discovery Into Health

Conclusion

- It is time to dig into the enormous collection of clinical data sleeping in the PACS.
- Three attempts are demonstrated to utilize retrospective clinical data for building large scale quality-labeled datasets for data-hungry learning paradigms.
- Discussed methods could be applied on many hospital PACS system to mine various collections of data.
- By mining and sharing all these existing data, the medical image analysis and diagnosis could be improved significantly using cutting-edge machine learning techniques.

NIH National Institutes of Health
Turning Discovery Into Health

Scan to contact

Thank you!

Those are joint works of many colleagues in NIH, including **Yifan Peng**, **Ke Yan**, **Le Lu**, Hoo-chang Shin, Zhiyong Lu, Mohammadhadi Bagheri, Lauren Kim, Isabella Nogues, Jianhua Yao and Ronald M. Summers.

Acknowledgement

- This work is supported by the Intramural Research Program of the National Institutes of Health Clinical Center and National Library of Medicine.
- This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov)
- We thank Nvidia corporation for the GPU donation.