VISVESWARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELGAUM-590014

TECHNICAL SEMINAR REPORT

On

“SEMINAR TITLE“

Submitted in the partial fulfillment of the requirements for V Semester

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By:

ABHILASH K R
V Semester, B.E, CSE

Under the guidance of:
Prof. SEMINAR GUIDE NAME
Designation
Dept of CSE
PESIT,Bangalore.

Carried out at:

PES INSTITUTE OF TECHNOLOGY

(an autonomous institute under VTU)

Department of Computer science & Engineering
100 Feet Ring Road , Banashankari III Stage,
Bangalore-560 085.

During the Academic Year
2016-2017

PES INSTITUTE OF TECHNOLOGY

(an autonomous institute under VTU)

100 Feet Ring Road, BSK 3rd Stage,Bangalore-560085

CERTIFICATE

This is to certify that the Technical Seminar entitled **"Seminar title",** bonafide work carried out by **Mr. ABHILASH K R** bearing USN **1PI14CS003** a student of **PES Institute of Technology (an autonomous institute under VTU)**in partial fulfillment for the award of **Bachelor of Engineering in Computer Science & Engineering** under **Visveswaraya Technological University, Belgaum** during the year 2016-17. The report has been approved as it satisfies the academic requirements in respect of Technical Seminar as prescribed for the said Degree.

…………………………..	………………………….
Signature of the Guide:	Signature of the HOD:
Prof. Seminar Guide Name	**Prof. NITIN V PUJARI**
Designation, Dept of CSE	HOD, Dept of CSE
PESIT, Bangalore	PESIT, Bangalore

Date: 4th October 2016

Place: Bengaluru

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this project would be incomplete without the mention of the people who made it possible. I consider myself privileged to express gratitude and respect towards all those who guided me through the completion of this project.

I would also like to express my gratitude to **Dr.K.S.Sridhar** (Principal, PESIT) **Prof. Nitin V Pujari**,(HOD ,Dept of CSE) for providing us congenial environment to work in and present the seminar.

I convey my sincere thanks to my seminar guide **Ms.Divya S J** (Asst.Professor, Dept of CSE,PESIT) for providing encouragement, constant support and guidance which was of a great help to complete this seminar successfully.

I am grateful to **Prof.Krupesha**, (Professor, Dept of CSE, PESIT) and **Prof. Arpitha** (Asst.Professor, Dept of CSE,PESIT) for giving me the support and guidance that was necessary for the completion of this seminar.

Abhilash K R

V SEM ,BE

ABSTRACT

Despite tremendous progress in computer vision, there has not been an attempt to apply machine learning on very large-scale medical image databases. Here they present an interleaved text/image deep learning system to extract and mine the semantic interactions of radiology images and reports from a national research hospital's Picture Archiving and Communication System. With natural language processing, we mine a collection of 216K representative two-dimensional images selected by clinicians for diagnostic reference and match the images with their descriptions in an automated manner. They then employ a weakly supervised approach using all of our available data to build models for generating approximate interpretations of patient images. Finally, they demonstrate a more strictly supervised approach to detect the presence and absence of a number of frequent disease types, providing more specific interpretations of patient scans. A relatively small amount of data is used for this part, due to the challenge in gathering quality labels from large raw text data. Our work shows the feasibility of large-scale learning and prediction in electronic patient records available in most modern clinical institutions.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER I

INTRODUCTION

In the medical domain, however, there are no similar large-scale labeled image data sets available. On the other hand, large collections of radiology images and reports are stored in many modern hospitals' Picture Archiving and Communication Systems (PACS). The invaluable semantic diagnostic knowledge inhabiting the mapping between hundreds of thousands of clinician-created high-quality text reports and linked image volumes remains largely unexplored. One the our primary goals is to extract and associate radiology images with clinically semantic labels via interleaved text/image data mining and deep learning on a large-scale PACS database (780K imaging examinations). To the best of knowledge, this is the first reported work performing automated mining and prediction on a hospital PACS database at a very large scale.

The Radiology reports are text documents describing patient history, symptoms, image observations and impressions written by board-certified radiologists. However, the reports do not contain specific image labels to be trained by a machine learning algorithm. Building the ImageNet database was mainly a manual process: harvesting images returned from Google image search engine according to the WordNet ontology hierarchy and pruning falsely tagged images using crowd-sourcing such as Amazon Mechanical Turk (AMT). This does not meet data collection and labeling needs due to the demanding difficulties of medical annotation tasks and the need for data privacy. Thus, they first propose to mine categorical semantic labels using a non-parametric topic modeling method latent Dirichlet Allocation (LDA) to provide a semantic interpretation of a patient image in three levels. While this provides a first-level interpretation of a patient image, labeling based on categorization can be nonspecific. To alleviate the issue of non-specificity, we further mine specific disease words in the reports mentioning the images. Feed-forward CNNs were then used to train and predict the presence/absence of the specific disease-categories.

CHAPTER II

LITERATURE SURVEY

Some of the papers or books that were referred by the authors of this paper are…

1. 'Latent Dirichlet Allocation'

   Authors : David M Blei, Andrew Y Ng, and Michael I Jordan

   This paper gives a detail explanation of one of the most popular topic modeling methods that is 'Latent Dirichlet Allocation'(LDA). A generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

2. 'Matching Words and Pictures'

   Authors : Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan

   An approach for modeling multi-modal data sets, focusing on the specific case of segmented images with associated text. Learning the joint distribution of image regions and words has many applications.

3. 'Natural language processing with Python'

   Authors : Steven Bird, Ewan Klein, and Edward Loper

   A book describing various algorithms and techniques related to NLP

CHAPTER III

Document Topic Learning with Latent Dirichlet Allocation

It is difficult to annotate all the images and the sentences referring to them. Unlike the images of ImageNet which often have a dominant object appearing in the center, our key images are mostly CT and MRI slices showing several organs usually with pathologies. There is a high amount of intrinsic ambiguity in defining and assigning a semantic label set to images, even for experienced clinicians. We therefore propose to mine image categorization labels using the non-parametric topic-modeling algorithm of radiology text reports in PACS.

There are some other popular methods for document topic modeling, such as Probabilistic Latent Semantic Analysis (pLSA) and Non-negative Matrix Factorization (NMF). In a study done LDA showed the most favorable results overall in human evaluations of the generated topics compared to other popular methods.

LDA offers a hierarchy of extracted topics and the number of topics can be chosen by evaluating each model's perplexity score (Equation 1), which is a common way to measure how well a probabilistic model generalizes by evaluating the log-likelihood of the model on a held-out validation set.

For an unseen document we set D_{val}, where M is the number of documents in the validation set, wd the words in the unseen document d, N_d the number of words in document d, with ϕ the topic matrix, and α the hyper-parameter for topic distribution of the documents.

$$perplexity(D_{val}) = \exp -\left\{ \sum_{d=1}^{M} \log p(wd \mid \phi,, \alpha) \Big/ \sum_{d=1}^{M} Nd \right\}$$

A lower perplexity score generally implies a better fit of the model for a given document Set.

The perplexity scores keep decreasing with an increasing number of topics; we choose the topic count to be 1000 as the rate of the perplexity score decrease is very small. (Figure 1).

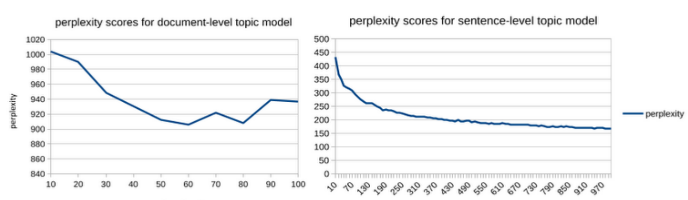


Figure 1

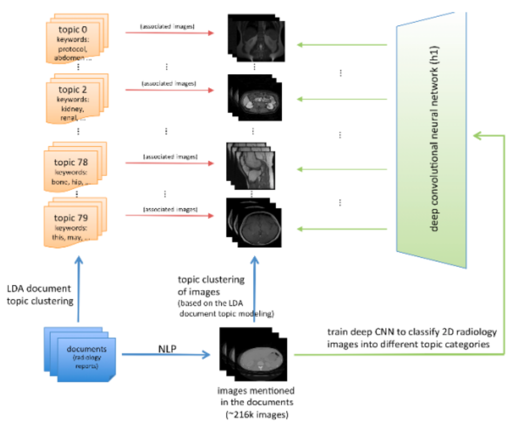
CHAPTER IV

Image to Document Topic Mapping with Deep Convolutional Neural Networks

For each level of topics discussed in Chapter 3, we train deep CNNs to map the images into document categories using the Caffe framework.We split our whole key image data set as follows: 85% used as the training data set, 5% as the validation, and 10% as the test data set. If a topic has too few images to be divided into training/validation/test for deep CNN learning, then that topic is neglected for the CNN training. Classifying images to

Document-level topics :

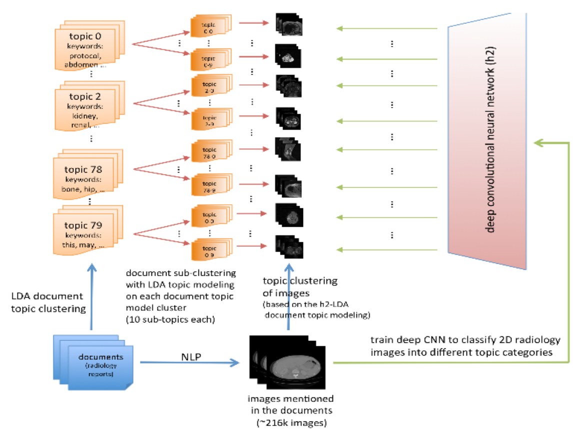
Figure 2



(a)

Document-level sub-topics :

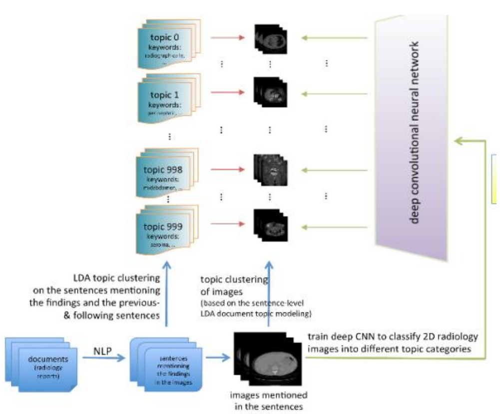
Figure 3



(b)

Sentence level-topics :

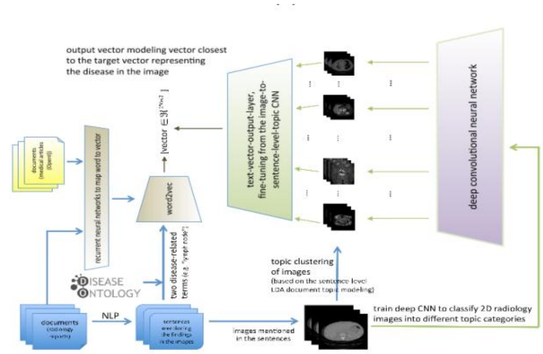
Figure 4



(c)

Image-word model :

Figure 5



(d)

CHAPTER IV

Generating Image-to-Text Description

The image-to-topic mapping in the previous chapter is a promising first step towards large-scale automated medical image interpretation. However, generating image descriptions will be more readily interpretable and descriptive. In addition, key words in the topics can help to understand the content of a given image with more semantic meaning. They, therefore propose to generate relevant key-word text descriptions using deep language/image CNN models.

Word-to-Vector Modeling :

Words with similar meaning are mapped or projected to closer locations in the vector space than dissimilar ones. An example visualization of the word vectors on a two-dimensional space using principal component analysis is shown in Figure 6.

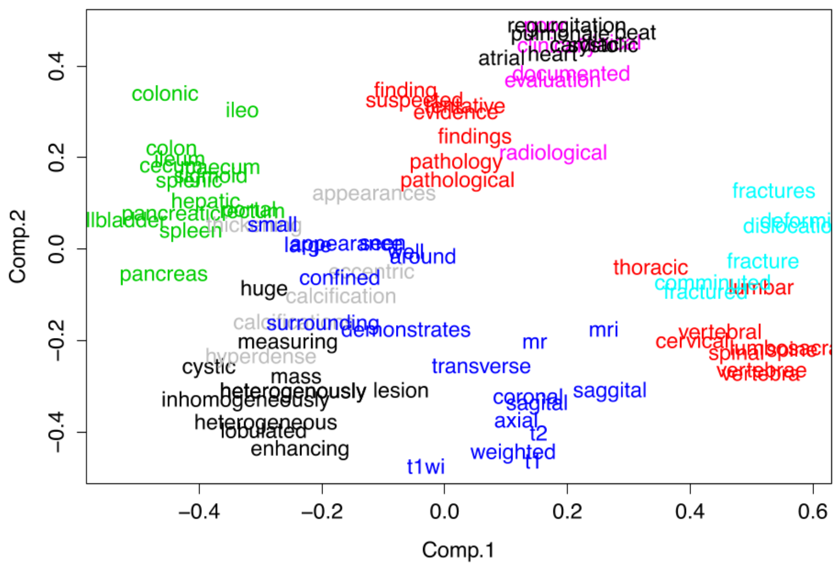


Figure 6

Some examples of query words and their corresponding closest words with respect to cosine similarity for the word-to-vector models which are trained on radiology reports only and with additional OpenI articles, are shown in Figure 7.

~1.2 billion words with OpenI		~1.2 billion words with OpenI		~1.2 billion words with OpenI		~1.2 billion words with OpenI	
"cyst"		**"heart"**		**"brain"**		**"liver"**	
cysts	0.799191	cardiac	0.672690	hemisphere	0.684149	hepatic	0.764163
hydatid	0.734686	respiratory	0.644453	hemispheric	0.668626	spleen	0.683242
cystic	0.701855	beat	0.642630	cerebellum	0.663902	cirrhotic	0.664428
unilocular	0.654273	pressure	0.558879	whole	0.661564	cirrhosis	0.664262
tailgut	0.639764	murmur	0.551323	regions	0.647632	hcc	0.656473
nonparasitic	0.621647	systolic	0.548490	mri	0.646674	portal	0.610437
epidermoid	0.604492	pericardial	0.538957	structural	0.638171	hepatocellular	0.603930
lipoma	0.588372	dobutamine	0.537429	neuroanatomical	0.636563	parenchyma	0.597169
cheesy	0.586947	intracardiac	0.533799	crinion	0.626951	splenic	0.579957
multiloculated	0.584199	great	0.532735	in	0.626707	hepatomegaly	0.573687
pearly	0.583126	rate	0.531352	parasaggital	0.618392	tumor	0.571135
multilocular	0.582670	beats	0.524729	illustration	0.610440	abdomen	0.559092
lesion	0.579009	atrial	0.524052	striatal	0.609282	hepatectomy	0.556156
tgdc	0.578533	tachycardia	0.521093	brains	0.607442	bclc	0.546798
multiseptate	0.575851	minute	0.520249	behavioral	0.606803	subcapsular	0.542745

~1 billion words reports only		~1 billion words reports only		~1 billion words reports only		~1 billion words reports only	
"cyst"		**"heart"**		**"brain"**		**"liver"**	
cysts	0.768382	lungs	0.526600	t1	0.615066	spleen	0.759884
septated	0.586067	mediastinum	0.517008	mri	0.595027	gallbladder	0.648075
polyp	0.583761	consolidating	0.486605	sagittal	0.580841	hepatomegaly	0.642022
simple	0.534717	pa	0.449816	flair	0.565445	gallstones	0.611837
septation	0.500951	chest	0.433362	t2	0.555053	pancreas	0.608356
parapelvic	0.500877	infiltrates	0.428404	axial	0.554040	gallstone	0.606063
incidental	0.500760	hyperinflated	0.413326	spgr	0.520954	steatosis	0.601081
small	0.487211	cardiomegaly	0.410785	weighted	0.502047	dome	0.594812
cystic	0.477632	hyperlucent	0.400836	technique	0.487768	portal	0.570008
pole	0.471933	pectus	0.396142	astrocytoma	0.480527	ascites	0.551869
multiseptated	0.469851	great	0.395712	gbm	0.476956	hepatosplenomegaly	0.540501
polyps	0.464380	ectatic	0.394560	gradient	0.476593	hepatic	0.537453
exophytic	0.459088	shifted	0.389205	oligodendroglioma	0.465892	cirrhosis	0.530389
hyperdense	0.457558	ray	0.389091	postcontrast	0.463686	fatty	0.522134
mucous	0.448427	infiltrate	0.387224	3d	0.458123	kidneys	0.515252

Figure 7

Image-to-Description Relation Mining and Matching :

The sentence referring to a key image and its adjacent sentences may contain a variety of words, but we are mostly interested in the disease-related terms which are highly correlated to diagnostic semantics. The sentences referring to an image and their adjacent sentences have 50.08 words on average; the number of disease-related terms in the three consecutive sentences is 5.17 on average with a standard deviation of 2.5. Therefore, we chose to use bi-grams for the image descriptions, to achieve a good trade-off between the medium level complexities without neglecting too many text-image pairs.

A deep regression CNN model is employed here, to map an image to a continuous output word-vector space from an image. The resulting bi-gram vector can be matched against a reference disease-related vocabulary in the word-vector space using cosine similarity.

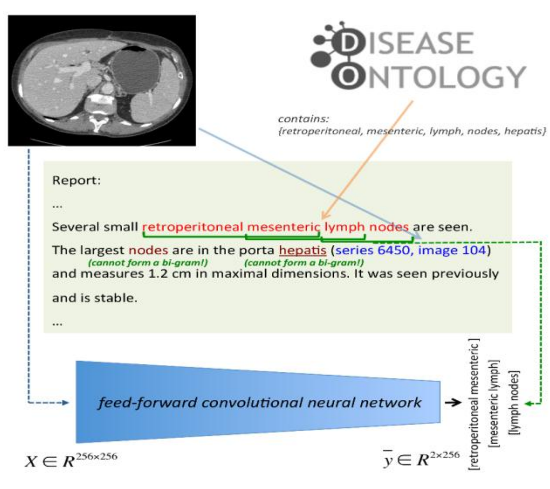


Figure 8

Key-Word Generation from Images :

For any key image in testing, first, we predict its topics at three levels (document-level, document-level sub-topics, sentence-level) using the three deep CNN models. Based on each word's probability of appearing in the LDA document topic, the fifty key-words with highest probability are mapped into the word-to-vector space of multivariate variables.

The closest key-words at three levels of topics (with the highest cosine similarity against either of the bi-gram words) are kept per image. Refer Figure 9 to see the results.

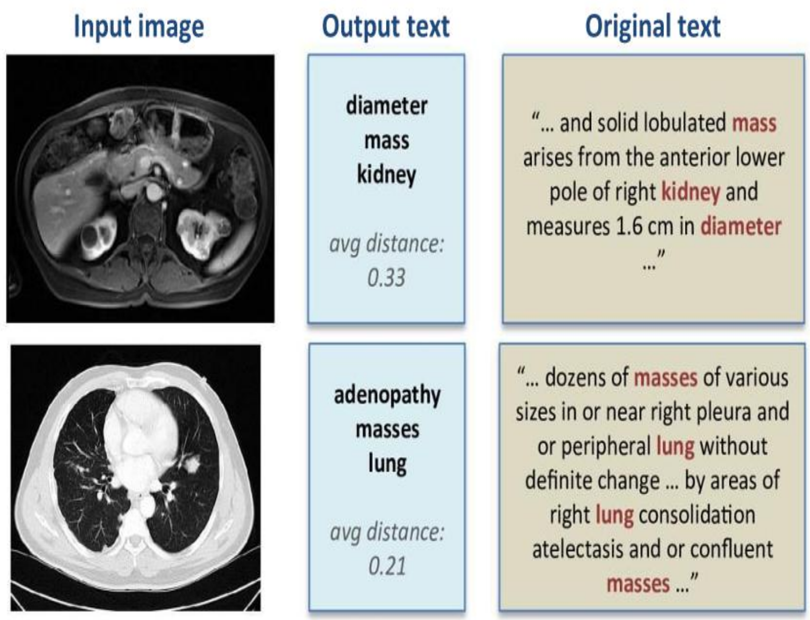


Figure 9

The rate of predicted disease-related words matching the actual words in the report sentences of test set is 0.56.

CHAPTER V

Predicting Presence or Absence of Frequent Disease Types

While the key-words generation in the previous section can aid the interpretation of a patient scan, the generated key-words, for example, spine, lung, are not very specific to a disease in an image. Nonetheless, one of the ultimate goals for large-scale radiology image/text analysis would be to automatically diagnose disease from a patient scan.

Mining Presence/Absence of Frequent Disease Terms :

The disease names in Disease Ontology (DO) contain not only disease terms but also non-disease terms describing a disease. The Unified Medical Language System (UMLS) integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more elective and inter-operable biomedical information systems and services, including electronic health records.

We are not only interested in disease terms associated with an image, but also whether the disease mentioned is present or absent. We use the assertion/negation detection algorithm to detect presence and absence of disease terms. The algorithm locates trigger terms which can indicate a clinical condition as negated or possible and determines which text falls within the scope of the trigger terms. The number of occurrences detected by assertion or negations in radiology reports are shown in Figure 10.

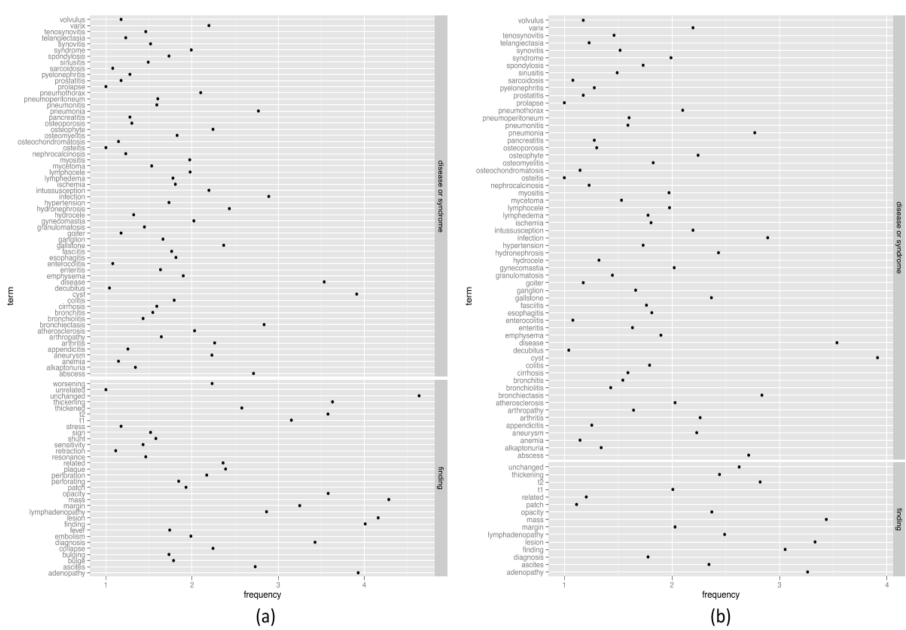


Figure 10

Predicting Disease in Images using CNN :

Similarly to the object detection task in the ImageNet challenge, we match and detect disease terms found in the sentences of radiology reports referring to an image using a CNN and softmax cost function. In addition to assigning disease terms to images, we also assign negated disease terms as the absence of the diseases in the images. If more than one disease term is mentioned for an image, we simply assign the terms multiple times for an image. Some statistics on the number of assertion/negation occurrences per image are shown in Table 1.

We fine-tune from the image to sentence-level-topic model, as the image-to-sentence-level-topic seems to be most closely related to the image-to-disease-specific-terms mode.

# images		per image mean/std		# assertions per image		# negations per image	
total matching	18291	# assertions mean	1.05	1/image	16133	1/image	1581
total not matching	197495	# negations mean	1.05	2/image	613	2/image	84
with assertions	16827	# assertions std	0.23	3/image	81	3/image	0
with negations	1665	# negations std	0.22	4/image	0	4/image	0

Table 1

CHAPTER VI

RESULTS AND VALIDATION

Though this method produces valid results but practically doctors can't rely on these as these results do not give the cause for the disease instead they give the specific terms related to the diseases present in the MRI or CT scan. The results can be seen in Figures 11 and 12.

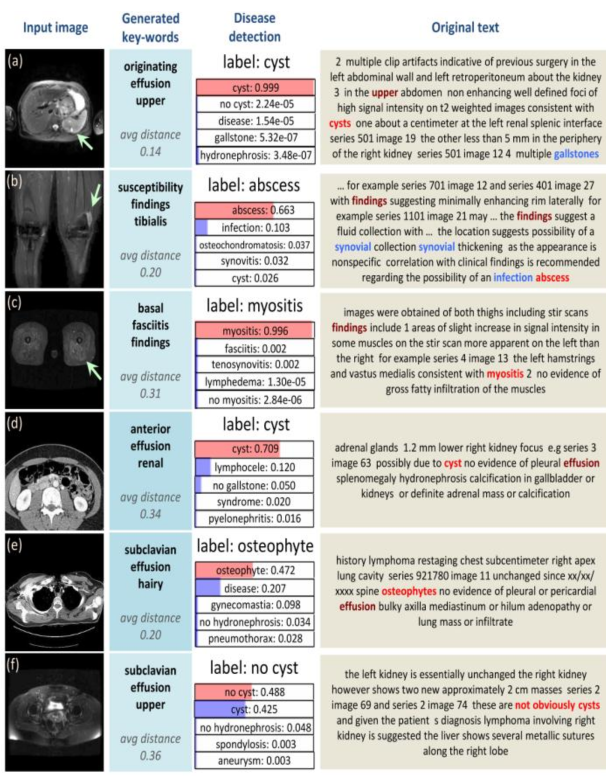


Figure 11

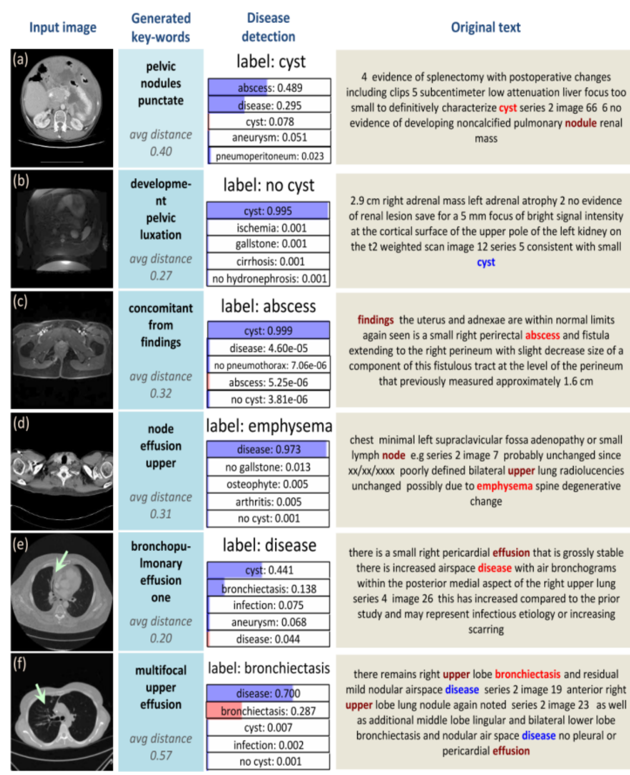


Figure 12

CONCLUSION

In this paper, they present an interleaved text/image deep mining system to extract the semantic interactions of radiology reports and diagnostic key images at a very large, unprecedented scale in the medical domain. Images are classified into hierarchies of topics according to their associated documents, and a neural language model is learned to assign disease terms to predict the image interpretation. However, by generating the "attributes" of patient images, the generated descriptions are not disease-specific, whereas one of the primary goals for medical image analysis is to automatically diagnose diseases. In order to address this issue, we mine and match frequent disease types using disease ontology and semantics, and demonstrate prediction of the presence/absence of disease with probability outputs. Yet, only about 10% of the entire data set could be used for this study due to the challenge of more precisely matching the disease words with semantics. This raises interesting questions regarding the trade-offs in designing a machine learning system analyzing large medical data.

REFERENCES

[1] Openi - an open access biomedical image search engine. http://openi.nlm.nih.gov. Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine.

[2] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. Journal of machine learning research

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In European Conference on Computer Vision, pages 663{676. Springer, 2010.

[4] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python. O'Reilly Media, Inc., 2009.

[5] David M Blei and Michael I Jordan. Modeling annotated data. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 127{134. ACM, 2003.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3:993{1022, 2003.

[7] Luke Carrivick, Sanjay Prabhu, Paul Goddard, and Jonathan Rossiter. Unsupervised learn- ing in radiology using novel latent variable models. In 2005 IEEE Computer Society Con- ference on Computer Vision and Pattern Recognition, volume 2, pages 854

[8] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated _ndings and diseases in discharge summaries. Journal of biomedical informatics, 34(5):301{310, 2001.

[9] Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise

[10] Text/Image Mining on a Large-Scale Radiology Database

Deleger. Extending the negex lexicon for multiple languages. Studies in health technology and informatics, 192:677, 2013.