Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database

Hoo-Chang Shin Le Lu Lauren Kim Ari Seff Jianhua Yao Ronald M. Summers

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory
Radiology and Imaging Sciences
National Institutes of Health Clinical Center
Bethesda, MD 20892-1182
{hoochang.shin, le.lu, lauren.kim2, ari.seff, rms}@nih.gov, jyao@cc.nih.gov

Abstract

Despite tremendous progress in computer vision, effective learning on very large-scale ($> 100K$ patients) medical image databases has been vastly hindered. We 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. Instead of using full 3D medical volumes, we focus on a collection of representative ~216K 2D key images/slices (selected by clinicians for diagnostic reference) with text-driven scalar and vector labels. Our system interleaves between unsupervised learning (e.g., latent Dirichlet allocation, recurrent neural net language models) on document- and sentence-level texts to generate semantic labels and supervised learning via deep convolutional neural networks (CNNs) to map from images to label spaces. Disease-related key words can be predicted for radiology images in a retrieval manner. We have demonstrated promising quantitative and qualitative results. The large-scale datasets of extracted key images and their categorization, embedded vector labels and sentence descriptions can be harnessed to alleviate the deep learning "data-hungry" obstacle in the medical domain.

1. Introduction

The ImageNet Large Scale Visual Recognition Challenge [8] provides more than one million labeled images from 1,000 object categories. The accessibility of a huge amount of well-annotated image data in computer vision rekindles deep convolutional neural networks (CNNs) [24, 41, 45] as a premier learning tool to solve the visual object class recognition tasks. Deep CNNs can perform significantly better than traditional shallow learning methods, but usually requires much more training data [24, 38]. In the medical domain, however, there are no similar large-scale labeled image datasets available. On the other hand, gigantic collections of radiology images and reports are stored in many modern hospitals' picture archiving and communication system (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 of our primary goals is to extract and associate radiology images with clinically semantic scalar and vector labels via interleaved text/image data mining and deep learning on a large-scale PACS database (~780K imaging examinations). To the best of our knowledge, this is the first reported work of "Deep Mining into PACS" at a very large scale.

Building the ImageNet database was mainly a manual process [8]: harvesting images returned from Google image search engine (according to the WordNet ontology hi-

total number of		# words in documents		# image modalities	
# documents	~780k	mean	131.13	CT	~169k
# images	~216k	std	95.72	MR	~46k
# words	~1 billion	max	1502	PET	67
# vocabulary	~29k	min	2	others	34

Table 1. Some statistics of the dataset. "Others" include CR (Computed Radiography), RF (Radio Fluoroscopy), and US (Ultrasound).

right	937k	images	312k	contrast	260k	unremarkable	195k
left	870k	seen	299k	axial	253k	lower	195k
impression	421k	mass	296k	lung	243k	upper	192k
evidence	352k	normal	278k	bone	219k	lesion	180k
findings	340k	small	275k	chest	208k	lobe	174k
CT	312k	noted	263k	MRI	204k	pleural	172k

Table 2. Examples of the most frequently occurring words in the radiology report documents.

erarchy) and pruning falsely tagged images using crowdsourcing such as Amazon Mechanical Turk (AMT). This does not meet our data collection and labeling needs due to the demanding difficulties of medical annotation tasks and the data privacy reasons. Thus we propose to mine image categorization labels from hierarchical, Bayesian document clustering method, e.g., generative latent Dirichlet allocation (LDA) topic modeling [6], using all available radiology text reports in PACS. 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 my a machine learning algorithm. We find that LDA-generated image categorization labels are valid, demonstrating good semantic coherence among clinician observers [22, 11], and can be effectively learned using deep CNNs with image inputs alone [24, 41]. Our deep CNN models on medical image modalities (mostly CT, MRI) are initialized with the model parameters pre-trained from ImageNet [8] using Caffe [19] framework. Kulkarni et al. [25] have spearheaded the efforts of learning the semantic connections between image contents and the sentences describing them (i.e., captions). Detecting objects of interest, attributes and prepositions and applying contextual regularization with a conditional random field (CRF) is a feasible approach [25] because many useful tools are available in computer vision. There has not yet been much comparable development on large-scale medical imaging understanding.

Our work has been inspired by the works building very large-scale image databases [8, 38] and the works establishing semantic connections of texts and images [25]. We observe good semantic coherence between labels obtained by hierarchical document topic models [6] and clinician's assessment. Based on this, both unsupervised (recurrent neural net language models [30, 32]) and supervised deep CNNs with categorization and regression losses are used for annotating large collection of radiology images. The fact that deep learning requires no hand-crafted image features is very desirable since significant adaption would be needed to apply conventional image features, e.g., HOG, SIFT for medical image learning. The large-scale datasets of extracted key images and their categorization, vector labels, describing sentences can be harnessed to alleviate deep learning's "data-hungry" challenge in the medical domain[1].

1.1. Related Work

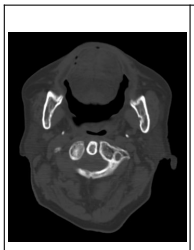
The ImageCLEF medical image annotation tasks of 2005-2007 have 9,000 training and 1,000 testing 2D images (converted as 32×32 pixel thumbnails in [9]) with 57 labels. Local image descriptors and intensity histograms are used in a bag-of-features approach in that work for this scene recognition-like problem. Unsupervised latent Dirichlet allocation based matching from lung disease words (e.g., fibrosis, emphysema) in radiology reports to 2D image blocks from axial CT chest scans (of 24 patients) is studied in [7]. This work is motivated by generative models of combining words and images [2, 5] under a very limited word/image vocabulary.

The most related works are [42, 11] which first map words into vector space using recurrent neural networks and then project images into the label-associated word-vector embeddings by minimizing the L_2 ([42]) or hinge rank losses ([11]) between the visual and label manifolds. The language model is trained on the texts of Wikipedia and tested on label-associated images from the CIFAR [23, 42] and ImageNet [8, 11] datsets. In comparison, our work is on a large, unlabeled medical dataset of associated images and text, where the text-derived labels are computed and verified with human intervention. Image-to-language correspondence was learned from ImageNet dataset and reasonably high quality image description datasets (Pascal1K [36], Flickr8K [16], Flickr30K [47]) in [20], where such caption datasets are not available in the medical domain. Graphical models have been employed to predict image attributes ([27, 39]), or to describe images ([25]) using manually annotated datasets ([36, 26]). Automatic label mining on large, unlabeled datasets is presented in [35, 18], however the variety of the label-space is limited (image text annotations). We analyze/mine the medical image semantics on both document and sentence levels, and deep CNNs are adapted to learn them from image contents [18, 41].

2. Data

To gain the most comprehensive understanding of diagnostic semantics, we use all available radiology reports of around ~780K imaging examinations, stored in the PACS of National Institutes of Health Clinical Center since the year 2000. Around $216K$ key 2D image slices (instead of all 3D image volumes) are studied here. Within 3D patient scans, most of the imaging information represented is normal anatomy, i.e, not the focus of the radiology reports. These "key images" were referenced (see Figure 1) by radiologists manually during radiology report writing, to provide a visual reference to pathologies or other notable findings. Therefore 2D key images are more correlated with the diagnostic semantics in the reports than the whole 3D scans, but not all reports have referenced key images ($215,786$ images from $61,845$ unique patients). Table 1 provides extracted database statistics, and Table 2 shows examples of the most frequently occurring words in radiology reports. Leveraging our deep learning models exploited in this paper will make it possible to automatically select key images from 3D patient scans to avoid mis-referencing.

[1]We are currently working on the institutional review board approval to share our extracted data (not original full radiology reports). We make our code and trained deep text/image models available in https://github.com/rsummers11/CADLab.



0001 REPORT : REASON FOR EXAM (Entered by ordering clinician into CRIS): hx of head and neck cancer. needs scan CT of the nasopharynx.
HISTORY: Head and neck cancer.
TECHNIQUE: Contiguous 2.5 mm axial images of the nasopharynx were performed without IV contrast. COMPARISON: xx/xx/xxxx.
FINDINGS: No soft tissue masses are seen within the soft tissues of the neck. The parotid and submandibular glands are predominantly fatty-replaced. Soft tissues of the Naso, oropharynx are unremarkable. There may be mild fat stranding of the right parapharyngeal soft tissues (series 1001, image 32). No abnormal masses are seen at that site. No bulky lymphadenopathy is seen. There is a fusiform aneurysm of the basilar artery as previously described. It appears to the mildly increased in size and currently measures 2.0 cm in transverse dimensions and previously measured 1.8 cm. It measured 1.5 cm in transverse dimensions on xx/xx/xxxx. Atherosclerotic calcifications are also seen within the carotid arteries bilaterally. There is near-complete opacification of the maxillary sinuses bilaterally. This has increased predominantly within the left maxillary sinus and mildly within the right maxillary sinus. The ethmoidal air cells are clear. Sphenoidal and frontal sinuses are clear. Degenerative changes of the cervical spine are noted.
IMPRESSIONS: 1. No soft tissue masses however, mild right parapharyngeal fat stranding is seen it may be postoperative or post radiation in nature. 2. Basilar artery aneurysm that has gradually increased in size when compared to prior examinations. 3. Atherosclerotic disease of the coronary arteries bilaterally..

0001 Report: CHEST, ABDOMEN, PELVIS CT: Multidetector helical (5 mm, quad) images following, and abdomen images prior to vascular contrast infusion (45 s delay, 2 cc/s, 130 cc Isovue) obtained without apparent complication. History: renal cell pt on Medarex protocol here for end of course evaluation.
CHEST: Multiple right, and at least one left lung masses minimally-moderately increasing since xx/xx/xxxx, compatible with metastases despite moderate decrease in at least one right mid-lung mass (e.g. series 4 image 30). Minimal pretracheal and subcarinal adenopathy increasing. Spine osteophytes. Enlargement thyroid on right side, and thyroid heterogeneity unchanged, possibly due to goiter. No evidence of pleural or pericardial effusion, axilla or left hilum adenopathy.
ABDOMEN, PELVIS: Few right and left liver foci, left periaortic and left adrenal fossa, and right sacrum mass and lytic lesion (series 3 image 88-95) increasing minimally, compatible with metastases. Scattered lumbar vertebra and bilateral ilium foci (e.g. series 3 image 55, 60, 80, 84-7, 96) foci possibly due to bone metastases. Uterine fundus focus (series 3 image 95) increasing in density since xx/xx/xxxx, possibly due to fibroid, metastasis. No evidence of splenomegaly, hydronephrosis, gallbladder calcification, or bulky mesenteric adenopathy.

Figure 1. Two examples of radiology reports and the referenced "key images" (providing a visual reference to pathologies or other notable findings).

Finding and extracting key images from radiology reports is done by natural language processing (NLP), i.e, finding a sentence mentioning a referenced image. For example, "*There may be mild fat stranding of the right parapharyngeal soft tissues (series 1001, image 32)*" is listed in Figure 1. The NLP steps [4] are sentence tokenization, word/number matching and stemming, and rule-based information extraction (e.g., translating "image 1013-78" to "images 1013-1078"). A total of ~187K images can be retrieved and matched in this manner, whereas the rest of ~28K key images are extracted according to their reference accession numbers in PACS. Our report-extracted key image database is the largest one ever reported and is highly representative of the huge collection of radiology diagnostic semantics over the last decade. Exploring effective deep learning models on this database opens new ways to parse and understand large-scale radiology image informatics.

3. Document Topic Learning with Latent Dirichlet Allocation

We propose to mine image categorization labels using unsupervised document topic-modeling algorithm, e.g. latent Dirichlet allocation (LDA) [6], on the ~780K radiology text reports in PACS. Unlike images from ImageNet [8, 38] which often have a dominant object appearing in the center, our key images are CT/MRI slices showing several coexisting organs/pathologies. There are high amounts of intrinsic ambiguity in defining and assigning a semantic label set to images, even for experienced clinicians. Our *hypothesis* is that the large collection of sub-million radiology reports statistically defines the categories meaningful for topic-mining (LDA) and visual learning (deep CNN).

LDA [6] was originally proposed to find latent topic models for a collection of text documents (e.g., newspapers). There are some other popular methods for document topic modeling, such as Probabilistic Latent Semantic Analysis (pLSA) [17] and Non-negative Matrix Factorization (NMF) [29]. We choose LDA for extracting latent topic labels among radiology report documents because LDA is shown to be more flexible yet learns more coherent topics over large sets of documents [43]. Furthermore, pLSA can be regarded as a special case of LDA [13] and NMF as a semi-equivalent model of pLSA [12, 10].

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 test set. For an unseen document set D_{test}, the perplexity score is defined as in Equation 1, where M is the number of documents in the test set (unseen hold-out set of documents), $\mathbf{w}_d$ the words in the unseen document d, N_d the number of words in document d, with $\mathbf{\Phi}$ the topic matrix, and α the hyperparameter for topic distribution of the documents.

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d|\mathbf{\Phi},\alpha)}{\sum_{d=1}^{M}N_d}\right\} \tag{1}$$

A lower perplexity score generally implies a better fit of the model for a given document set [6].

Based on the perplexity score evaluated on 80% of the total documents used for training and 20% used for testing, the number of topics chosen is 80 for the document-level model using perplexity scores for model selection (Figure 2). Although the document distribution in the topic space is approximately balanced[2], the distribution of image counts for the topics is unbalanced. Specifically, topic #77 (non-primary metastasis spreading across a variety of body parts) contains nearly half of the ~216K key images. To address the data bias, a second-hierarchy topics are obtained for each of the first document-level topics, resulting in 800 topics, where the number of second-hierarchy topics is also chosen based on the average perplexity scores evaluated on each document-level topic. Lastly, to compare the method of using the whole report with using only the sentence directly describing the key images for latent topic mining, a sentence-level (third-hierarchy) LDA topics are obtained based on three sentences only: the sentence mentioning the key-image (Figure 1) and its adjacent sentences as proximal context. 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 beyond that point (Figure 2).

[2]Please refer to the supplementary material for the distribution of documents and images for LDA topics.

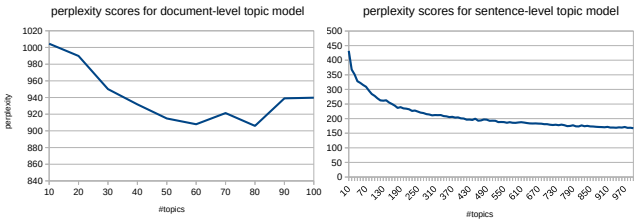


Figure 2. Perplexity scores for document-/sentence- level topic models. Number of topics with low perplexity score is selected as the optimal (80 for document-level, 1000 for sentence-level).

We observe that LDA-generated image categorization labels are valid, demonstrating good semantic coherence among clinician observers [22, 11]. The lists of key words and sampled images per topic label are subjected to board-certified radiologist's review and validation. Some examples of document-level topics with their corresponding images and topic key words are shown in Figure 3. Based on radiologists' review, our LDA topics discover semantics at different levels: 73 low-level concepts (e.g., pathology examination of certain body regions and organs: topic #47 - sinus diseases; #2 - lesions of solid abdominal organs, primarily kidney; #10 - pulmonary diseases; #13 - brain MRI; #19 - renal diseases on mixed imaging modalities; #36 - brain tumors). There are 7 mid- to high-level concepts (e.g., topic #77 - non-primary metastasis spreading across a variety of body parts; topic #79 - cases with high diagnosis uncertainty/equivocation; #72 - indeterminate lesions; #74 - instrumentation artifacts limiting interpretation). Low-level topic images are visually more coherent (i.e., may be easier to learn). High-level topics may be analogous to [21, 34]. About half of the key images are associated with topic #77, implying that the clinicians' image referencing behavior patterns heavily focuses on metastatic patients. For more details and the image-topic associations, refer to Figure 3, Figure 4 and supplementary material. Even though LDA labels are computed with text information only, we next investigate the plausibility of mapping images to the topic labels (at all semantic levels) via deep CNN models.

4. Image to Document Topic Mapping with Deep Convolutional Neural Networks

For each level of topics discussed in Section 3, we train deep CNNs to map the images into document categories using Caffe [19] framework. While the images of some topic categories and some body parts are easily distinguishable (e.g. Figure 3), the visual differences in abdominal parts are rather subtle (e.g. Figure 4). Distinguishing the subtleties and high-level concept categories in the images could benefit from a more complex model so that the model can handle these subtleties.

We split our whole key image dataset as follows: 85% used as the training dataset, 5% as the cross-validation (CV) and 10% as the test dataset. If a topic has too few images to be divided into training/CV/test for deep CNN learning (normally rare imaging protocols), then that topic is neglected for the CNN training (e.g., topic #5 Abdominal ultrasound, #28, #49 DEXA scans of different usages). In total, 60 topics were used for the document-level image-topic mapping, 385 for the second-hierarchy document-level mapping, and 717 for the sentence-level mapping. Surprisingly, we find that transfer learning from the ImageNet pre-trained CNN parameters on natural images to our medical image modalities (mostly CT, MRI) significantly helps the image classification performance[3]. Thus our CNN models are fined-tuned from the ImageNet CNN models by default. Similar findings of the deep feature generality across different image modalities have been reported [15, 14] but are empirically verified with only much smaller datasets than ours. Our key image dataset is ~1/5 size of ImageNet [38] as the largest annotated medical image dataset to date.

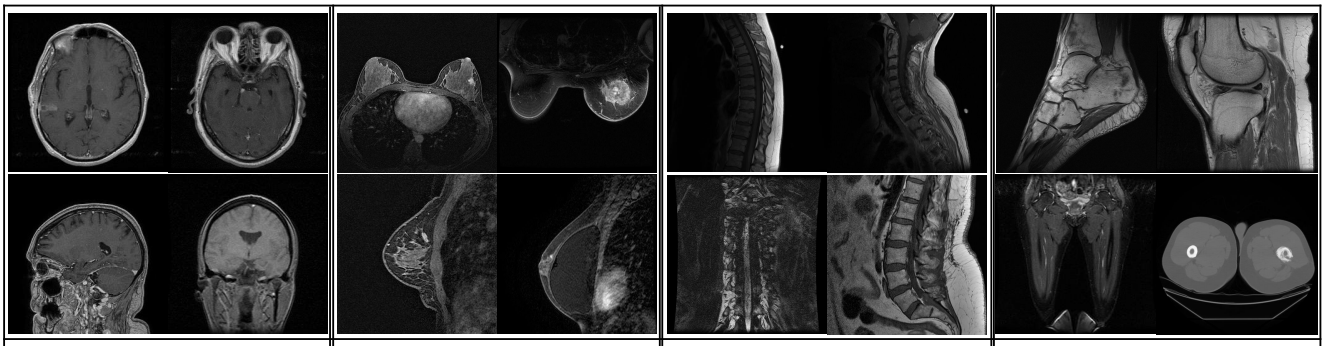
Implementation & Results: All our CNN network settings are similar[4] or same as the ImageNet Challenge "AlexNet" [24] and "VGG-19" [41] models. For image categorization, we change the numbers of output nodes in the last softmax classification layer, i.e., 60, 385 and 717 for the document-level, document-level-h2, and sentence-level respectively. The networks for first-level semantic labels are fine-tuned from the pre-trained ImageNet models, where the networks for the lower-level semantic labels are fine-tuned from the models of the higher-level semantic labels [5]. For all the CNN layers except the newly modified ones, the learning rate is set 0.001 for weights and biases, momentum 0.9, weight decay 0.0005 and a smaller batch size 50 (as opposed to 256 [41]). These adapted layers are initialized from random and their learning rates are set higher – learning rate: 0.01 for weight; 0.02 for bias; weight decay: 1 for weight; 0 for bias. All the key images are resampled to the spatial resolution of 256×256 pixels, mostly from the original 512×512. Then we follow [41] to crop the input images from 256×256 to 227×227 for training.

We would expect that the level of difficulties for learning and classifying the images into the LDA-induced topics will be different for each semantic level. Low-level semantic classes can have key images of axial/sagittal/coronal slices with position variations and across MRI/CT modalities. Some body parts and topics, e.g., #63 pelvic (female reproductive tract) imaging, are visually more challenging

[3] Please refer to the supplementary material for the details of the experiments.

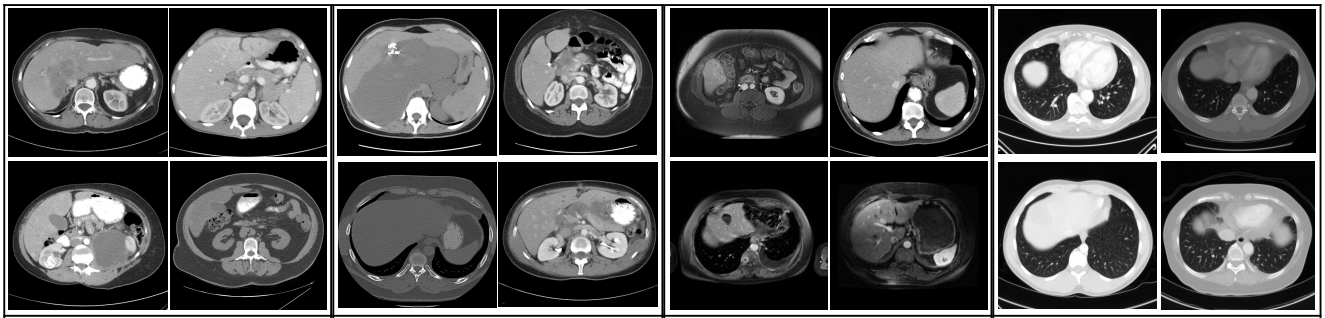
[4] We used Caffe reference network [19] which is a slight modification to the "AlexNet" [24].

[5] Detailed experiments showing benefits of transferring the parameters from the related tasks can be found in the supplementary material.



Topic 04:	Topic 17:	Topic 31:	Topic 78:
axial,contrast,mri,sagittal,post,flair,enhancement,blood,dynamic,brain,relative,volume,this,precontrast,from,tesla,fse,diffusion,gradient,resection,comparisons,maps,philips,progression,some,susceptibility,perfusion,stable,achieva,technique,echo,weighted,1.5,evidence,mass,findings,hemorrhage,enhanced,impression,frontal,signal,coronal,dti,tumor,t1-ffe,hydrocephalus,magnevist,reformations,bolus,lesion	breast,performed,suspicious,breasts,seen,impression,mass,screening,mammogram,dated,annual,cancer,mri,benign,bilateral,was,bi-rads,mammograms,Negative,dense,history,calcifications,images,views,studies,quadrant,mammography,volume,organ,aspect,suggested,category,mastectomy,before,tissue,enhancement,microcalcifications,heterogeneously,prior,family,examination,recommend,malignancy,high,suggest,outer,masses,developing,clip,patient	spine,cord,cervical,thoracic,spinal,level,canal,lumbar,sagittal,vertebral,neural,disc,signal,mri,body,technique,levels,findings,foramina,mild,disk,nerve,within,small,marrow,central,bodies,normal,impression,enhancing,conus,syrinx,this,narrowing,lesions,roots,contrast,throughout,bone,degenerative,foramen,protrusion,multiple,l5-s1,also,abnormal,c5-c6,posterior,changes,heights	bone,lesion,hip,knee,femoral,lytic,femur,proximal,head,sclerotic,joint,shoulder,hips,evidence,pelvis,distal,lesions,findings,humeral,lateral,fracture,medial,humerus,focal,impression,bony,prosthesis,history,iliac,pain,bilateral,blastic,avn,acetabulum,seen,marrow,sclerosis,view,both,osteolytic,cortical,heads,area,cortex,effusion,replacement,tibial,involving,consistent,views

Figure 3. Examples of LDA generated document-level topics with corresponding images and key words. Topic #4 MRI of brain tumor; #17 breast imaging; #31 degenerative spine disc disease; #78 bone metastases.



Topic 77-0:	Topic 77-2:	Topic 77-5:	Topic 77-9:
kidney,images,abdomen,e.g,prior,mass,pancreas,following,cysts,adrenal,liver,foci,renal,contrast,approximate,including,focus,cyst,bilateral,masses,size,enhancing,for,also,given,possibly,mid,2.5,vascular,without,due,nephrectomy,please,1.5,from,few,multiphase,subcentimeter,least,comparison,patient,dual-phase,length,apparent,complication,obtained,upper,study,lower,vhl	bulky,pelvis,bone,gross,since,liver,abdomen,calcification,vascular,study,lung,mass,isovue,dfov,without,contrast,administration,impression,metastasis,chest,for,images,mesenteric,axilla,following,hilum,cc/s,helical,multidetector,ascites,enteric,reason,apparent,complication,pleural,splenomegaly,pericardial,hydronephrosis,delay,effusion,mediastinum,obtained,300,spine,gallbladder,report,130,retroperitoneal,spleen,e.g	images,axial,t1-weighted,without,prior,liver,following,t2-weighted,tesla,fat-suppressed,multiple,sequences,e.g,characteristic,obtained,1.5,foci,fat,abdomen,for,prolonged,coronal,including,relaxation,hydronephrosis,mri,magnevist,splenomegaly,complication,apparent,vascular,pleural,impression,report,effusion,contrast,reason,study,mass,administration,since,focus,multiphase,definite,echo,defect,gross,filling,ascites,into	lung,chest,pleural,images,bilateral,minimal,effusion,lower,obtained,pericardial,multidetector,helical,axilla,study,report,mass,infiltrate,for,scarring,since,bulky,and/or,clinical,splenomegaly,dfov,cavity,e.g,impression,decreasing,infiltrates,focal,mediastinum,disease,atelectasis,hydronephrosis,small,reason,upper,untoward,history,probable,appearing,calcification,lobe,8-channel,supine,scattered,prone,bone,intervals

Document-level Topic 77:
compatible,adenopathy,series,unchanged,image,evidence,images,e.g,pelvis,lung,since,abdomen,vascular,minimal,foci,bulky,mass,calcification,bone,chest,contrast,liver,effusion,pleural,obtained,gross,following,without,splenomegaly,axilla,hydronephrosis,metastasis,bilateral,pericardial,increasing,helical,multidetector,apparent,complication,hilum,due,spine,gallbladder,administration,mesenteric,fat,dfov,cc/s,appearing,delay

Figure 4. Examples of some second-hierarchy topics of document-level topic #77, with corresponding images and topic key-words. The key-words and the images for the document-level topic (#77) indicates metastatic disease. The key-words for topic #77 are: [abdomen,pelvis,chest,contrast,performed,oral,was,present,masses,stable,intravenous,adenopathy,liver,retroperitoneal,comparison,administration,scans,130,small,parenchymal,mediastinal,dated,after,which,evidence,were,pulmonary,made,adrenal,prior,pelvic,without,cysts,spleen,mass,disease,multiple,isovue-300,obtained,areas,consistent,nodules,changes,pleural,lesions,following,abdominal,that,hilar,axillary].

	AlexNet 8-layers			VGG 19-layers		
	CV	top-1	top-5	CV	top-1	top-5
document-level	0.6078	0.6072	0.9294	0.6634	0.6582	0.9460
document-level-h2	0.3448	0.3252	0.5632	0.5408	0.5390	0.6960
sentence-level	0.48	0.48	0.56	0.51	0.50	0.58

Table 3. Validation and top-1, top-5 test scores in classification accuracy using AlexNet [24] and VGG-19 [41] deep CNN models.

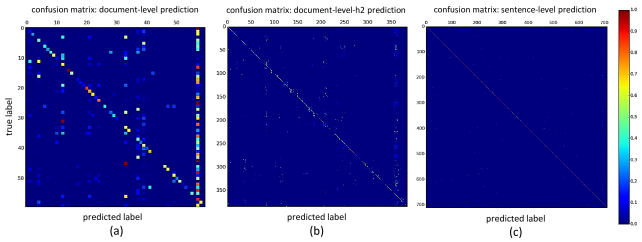


Figure 5. Confusion matrices of (a) document-, (b) second-hierarchy document-, (c) sentence- level classification [41] ((b) and (c) can be viewed best in electronic version of this document).

than others. Mid- to high-level concepts all demonstrate much larger within-class variations in their visual appearance since they are diseases occurring within different organs and are only coherent at high level semantics. Table 3 provides the validation and top-1, top-5 testing in classification accuracies for each level of topic models using AlexNet [24] and VGG-19 [41] based deep CNN models. Out of the three tasks, document-level-h2 is the hardest with document-level being relatively the easiest. Our top-1 testing accuracies are closely comparable with the validation ones, showing good training/testing generality and no observable over-fitting. All top-5 accuracy scores are significantly higher than top-1 values (increasing from 0.658 to 0.946 using VGG-19, or 0.607 to 0.929 via AlexNet in document-level), which indicates the classification errors or fusions are not uniformly distributed among other false classes. Latent "blocky subspace of classes" may exist (i.e., several topic classes form a tightly correlated subgroup) in our discovered label space, where the confusion matrices in Figure 5 verify this finding.

It is shown that the deeper 19-layer model (VGG-19 [41]) performs consistently better than the 8-layer model (AlexNet [24]) in classification accuracy, especially for document-level-h2. Compared with the ImageNet 2014 results, top-1 error rates are moderately higher (34% versus 30%) and top-5 test errors 6%~8% are comparable. In summary, our quantitative results are very encouraging given fewer image categorization labels (60 versus 1000) but much higher annotation uncertainties because of the unsupervised LDA topic models. Multi-level semantic concepts show good image learnability by deep CNN models which sheds light on the feasibility of automatically parsing very large-scale radiology image databases.

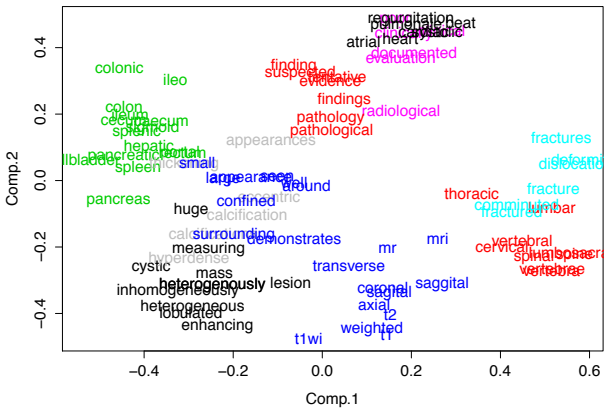


Figure 6. Example words embedded in the vector space using word-to-vector modeling (https://code.google.com/p/word2vec/) visualized on 2D space, showing (clinical) words with similar meanings are located nearby in the vector space.

5. Generating Image-to-Text Description

The deep CNN image categorization on multi-level document topic labels in Section 4 demonstrates promising results. The ontology of document clustering-discovered categories needs to be further reviewed and refined through a "clinician in-the-loop" process. In order to help understand the semantic contents of a given image in more detail, we propose to generate relevant key-word text descriptions [25] using deep language/image learning models.

5.1. Word-to-Vector Modeling and Removing Word-Level Ambiguity

In radiology reports, there exist many recurring word morphisms in text identification, e.g., [mr, mri, t1-/t2-weighted[6]], [cyst, cystic, cysts], [tumor, tumour, tumors, metastasis, metastatic], etc. We train a deep word-to-vector model [33, 32, 30] to address this word-level labeling space ambiguity. A total of ~1.2 billion words from our radiology reports as well as from biomedical research articles obtained from OpenI [1] are used. Words with similar meaning are mapped or projected to closer locations in the vector space than dissimilar ones (i.e., locality-preserving mapping). An example visualization of the word vectors on the 2-D space using PCA is shown in Figure 6.

A skip-gram model [30, 32] is employed with the mapping vector dimension of $\mathbb{R}^{256}$ per word, trained using the *hierarchical softmax* cost function, sliding-window size of 10 and frequent words sub-sampled in 0.01% frequency. It is found that combining an additional (more diverse) set of related documents, such as OpenI biomedical research articles, is helpful for the model to learn a better vector representation while keeping all the hyperparameters the same.

[6]Natural language expressions for imaging modalities of magnetic resonance imaging (MRI).

#words/sentence	mean	median	std	max	min
reports-wide	11.7	9	8.97	1014	1
image references	23.22	19	16.99	221	4
image references, no stopwords no digits	13.46	11	9.94	143	2
image references, disease terms only	5.17	4	2.52	25	1

Table 4. Some statistics about number of words per sentence – across the radiology reports (reports-wide), across the sentences referring the "key images" and its two adjacent ones (image references) and these not counting stopwords and digits as well as counting disease related words only.

Some examples of query words and their corresponding closest words in terms of cosine similarity for the word-vector models [33] trained on radiology reports only (total of ~1 billion words) and with additional OpenI articles (total of ~1.2 billion words) can be found in the supplementary material.

5.2. Image-to-Description Relation Mining and Matching

The sentence (and adjacent sentences) referring to a key image may contain a variety of words, but we are mostly interested in the disease-related terms (which are highly correlated to diagnostic semantics). To obtain only the disease-related terms, we exploit the human disease terms and their synonyms from the Disease-Ontology (DO) [40], a collection of 8,707 unique disease-related terms. While 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 complexity and not neglecting too many text-image pairs. Complete statistics about the number of words in the documents are shown in Table 4.

Bi-gram disease terms are extracted so that we can train a deep CNN (in Section 5.3) to predict the vector/word-level image representation ($\mathbb{R}^{256\times 2}$). If multiple bi-grams can be extracted per image from the sentence referring the image and the two adjacent ones, the image is trained as many times as the number of bi-grams with different target vectors ($\mathbb{R}^{256\times 2}$). If a disease term cannot form a bi-gram, then the term is ignored. This is a challenging *weakly annotated learning* problem using referring sentences for labels. This process is shown in Figure 7, and the illustrations for the complete work-flow can be found in the supplementary material. The bi-grams of DO disease-related terms in the vector representation ($\mathbb{R}^{256\times 2}$) are analogous to detecting multiple objects of interest and describing their spatial configurations in the image caption [25]. 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

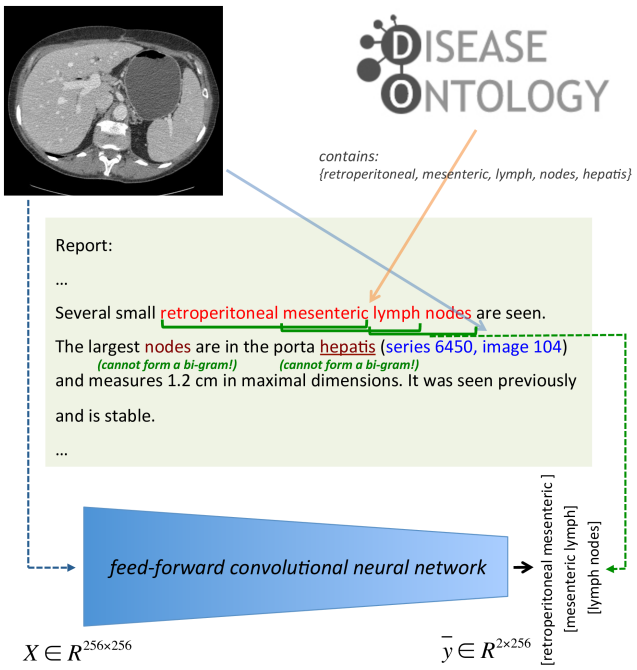


Figure 7. An example illustration of how word sequences are learned for an image. Bi-grams are selected from the image's reference sentences containing disease-related terms from the disease ontology (DO) [40]. Each bi-gram is converted to a vector of $\mathbf{Z} \in \mathbb{R}^{256\times 2}$ to learn from an image. Image input vectors as $\{\mathbf{X} \in \mathbb{R}^{256\times 256}\}$ are learned through a CNN by minimizing the cross-entropy loss between the target vector and output vector. The words "nodes" and "hepatis" in the second line are DO terms but are ignored since they can not form a bi-gram. This figure was reproduced with permission to use the DO logo from http://disease-ontology.org/.

disease-related vocabulary in the word-vector space using cosine similarity.

5.3. Image-to-Words Deep CNN Regression

It has been shown [44] that deep recurrent neural networks (RNN[7]) can learn the language representation for machine translation. To learn the image-to-text representation, we map the images to the vectors of word sequences describing the image. This can be formulated as a regression CNN, replacing the softmax cost in Section 4 with the cross-entropy cost function for the last output layer of VGG-19 CNN model [41]:

$$E = -\frac{1}{n}\sum_{n=1}^{N}[g(\mathbf{z})_n \hat{g}(\bar{\mathbf{z}}_n) + (1 - g(\mathbf{z}_n))\log(1 - g(\hat{\mathbf{z}}_n))], \tag{2}$$

where $\mathbf{z}_n$ or $\hat{\mathbf{z}}_n$ is any uni-element of the target word vectors $\mathbf{Z}_n$ or optimized output vectors $\hat{\mathbf{Z}}_n$, $g(x)$ is the sigmoid

[7]While RNN [37, 46] is the popular choice for learning language models [3, 31], deep CNN [28, 24] is more suitable for image classification.

function ($g(x) = 1/(1 + e^x)$), and n is the number of samples in the database.

We adopt the CNN model of [41] for the image-to-text representation since it works consistently better than the other relatively simpler model [24] in our image categorization task. We fine-tune the parameters of the CNNs for predicting the topic-level labels in Section 4 with the modified cost function, to model the image-to-text representation instead of classifying images into categories. The newly modified output layer has 512 nodes for bi-grams as 256×2 (double the dimensionality of the word vectors), with the cross-entropy cost decreasing and converging during training in about 10,000 iterations.

5.4. Word Prediction from Images as Retrieval

For any key image in testing, first we predict its categorization topic labels for each hierarchy (document-level, document-level-h2, sentence-level) using the three deep CNN models [41] in Section 4. Top 50 key-words in each LDA document-topics are mapped into the word-to-vector space as multivariate variables $\mathbb{R}^{256}$ (Section 5.1). Then, the image is mapped to a $\mathbb{R}^{256\times2}$ output vector using the bi-gram CNN model in Section 5.3. Lastly, we match each of the 50 topic key-word vectors ($\mathbb{R}^{256}$) against the first and second half of the $\mathbb{R}^{256\times2}$ output vector (i.e., treated as two words in the word-to-word matching) using cosine similarity.

The closest key-words at three hierarchy levels (with the highest cosine similarity against either of the bi-gram words) are kept per image. The rate of predicted disease-related words matching the actual words in the report sentences (recall-at-K, K=1 (R@1 score)) was 0.56. Text generation examples are shown in Figure 8, with three key-words from three categorization levels per image. We only report R@1 score on disease-related words compared to the previous works [20, 11] where they report from R@1 up to R@20 on the entire image caption words (e.g. R@1=0.16 on Flickr30K dataset [20]). As we used NLP to parse and extract image-describing sentences from the whole radiology reports, our ground-truth image-to-text associations are much noisier than the caption dataset used in [11, 20]. Also for that reason, our generated image-to-text associations are not as exact as the generated descriptions in [11, 20].

6. Conclusion & Discussion

It has been unclear how to extend the significant success in image classification using deep convolutional neural networks from computer vision to medical imaging. Open questions remain such as defining clinically relevant image labels, how to annotate the huge amount of medical images required by deep learning models, and to what extent and scale the deep CNN architecture is generalizable in medical image analysis.

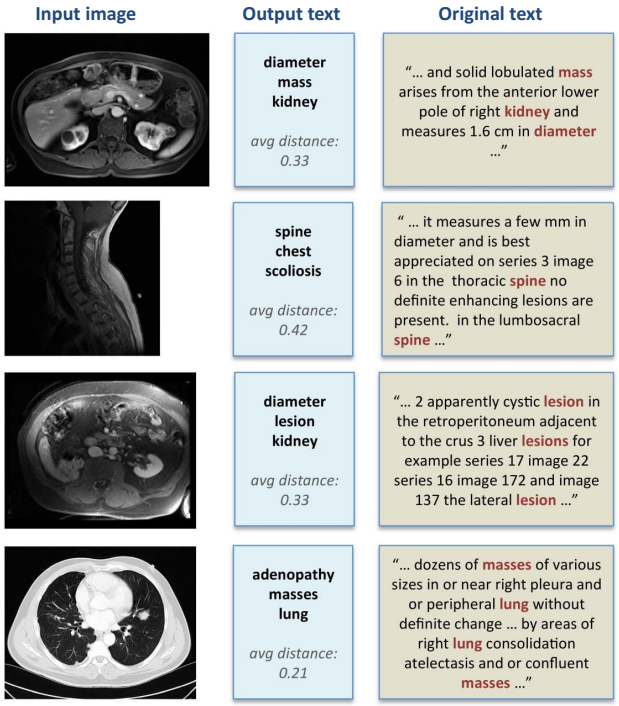


Figure 8. Examples of text key-word generation results, and average cosine distances between the generated words from the disease-related words in the original texts. It is also noticeable that kidney and adenopathy appear in the 3rd and 4th row images but were not mentioned in the reports. The rate of predicted disease-related words matching the actual words in the report sentences (recall-at-K, K=1 (R@1 score)) was 0.56.

In this paper, we present an interleaved text/image deep mining system to extract the semantic interactions of radiology reports and diagnostic key images at a very large and unprecedented scale in the medical domain. Images are classified into different levels of semantic topics according to their associated documents, and a neural language model is learned to assign field-specific (disease) terms to predict what is in the radiology image. We demonstrate promising quantitative and qualitative results, suggesting a way to extend the deep image classification systems to learning medical imaging informatics from "big-data" at a modern hospital scale.

To the best of our knowledge, this is the first study performing a large-scale image/text analysis on a hospital picture archiving and communication system (PACS) database. We hope that this study will inspire and encourage other institutions in mining other large unannotated clinical databases, to achieve towards establishing a central training resource and performance benchmark for large-scale medical image research, similarly to the ImageNet [8, 38] for computer vision.

Acknowledgments

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

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. 6

[2] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMRL*, 3:1107–1135, 2003. 2

[3] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006. 7

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009. 3

[5] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003. 2

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003. 2, 3

[7] L. Carrivick, S. Prabhu, P. Goddard, and J. Rossiter. Unsupervised learning in radiology using novel latent variable models. In *CVPR*, 2005. 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1, 2, 3, 8

[9] T. Deselaers and H. Ney. Deformations, patches, and discriminative models for automatic annotation of medical radiographs. *PRL*, 2008. 2

[10] C. Ding, T. Li, and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 342. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. 3

[11] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 2, 4, 8

[12] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602. ACM, 2005. 3

[13] M. Girolami and A. Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434. ACM, 2003. 3

[14] A. Gupta, M. Ayhan, and A. Maida. Natural image bases to represent neuroimaging data. In *ICML*, 2013. 4

[15] S. Gupta, R. Girshick, P. Arbelez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014. 4

[16] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013. 2

[17] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999. 3

[18] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *ECCV*, pages 512–528. 2014. 2

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 2, 4

[20] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2014. 2, 8

[21] H. Kiapour, K. Yamaguchi, A. Berg, and T. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 4

[22] R. Kiros and C. Szepesvri. Deep representations and codes for image auto-annotation. In *NIPS*, pages 917–925, 2012. 2, 4

[23] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 2009. 2

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2, 4, 6, 7, 8

[25] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2891–2903, 2013. 2, 6, 7

[26] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009. 2

[27] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 2

[28] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–97. IEEE, 2004. 7

[29] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 3

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 6

[31] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010. 7

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. 2, 6

[33] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013. 6, 7

[34] V. Ordonez and T. Berg. Learning high-level judgments of urban perception. In *ECCV*, 2014. 4

[35] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011. 2

[36] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010. 2

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1988. 7

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. 1, 2, 3, 4, 8

[39] W. Scheirer, N. Kumar, P. Belhumeur, and T. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012. 2

[40] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012. 7

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 4, 6, 7, 8

[42] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013. 2

[43] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012. 3

[44] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 2014. 7

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1

[46] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990. 7

[47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2