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 network 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.

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

<table>
<thead>
<tr>
<th>total number of</th>
<th># words in documents</th>
<th># image modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td># documents</td>
<td>~780K</td>
<td>mean 131.13 CT ~169K</td>
</tr>
<tr>
<td># images</td>
<td>~216K</td>
<td>std 95.72 MR ~46K</td>
</tr>
<tr>
<td># words</td>
<td>~1 billion</td>
<td>max 1502 PET 67</td>
</tr>
<tr>
<td># vocabulary</td>
<td>~29k</td>
<td>min 2 others 34</td>
</tr>
</tbody>
</table>

| right | 937k | images | 312k | contrast | 260k | unremarkable | 195k |
| left  | 870k | seen   | 269k | axial    | 253k | lower        | 195k |
|       |      | mass   | 290k | long     | 243k | upper        | 192k |
| evidence | 421k | CT     | 312k | noted    | 263k | lobe         | 174k |
| findings | 340k | MRI    | 204k | small    | 275k | pleural      | 172k |

Table 2. Examples of the most frequently occurring words in the radiology report documents.
Local image descriptors and intensity histograms are used to extract key images and their categorization, vectorization with a conditional random field (CRF) is a feasible approach because many useful tools are available in computer vision. There has not yet been much comparable work 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].

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 $32 \times 32$ 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] datasets. 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.
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), \( w_d \) the words in the unseen document \( d \), \( N_d \) the number of words in document \( d \), with \( \Phi \) the topic matrix, and \( \alpha \) the hyperparameter for topic distribution of the documents.

\[
\text{perplexity}(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d|\Phi, \alpha)}{\sum_{d=1}^{M} N_d} \right\}
\]  

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.

### 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
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 \(256 \times 256\) pixels, mostly from the original \(512 \times 512\). Then we follow \([41]\) to crop the input images from \(256 \times 256\) to \(227 \times 227\) 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

\(^{4}\)Please refer to the supplementary material for the details of the experiments.

\(^{5}\)We used Caffe reference network \([19]\) which is a slight modification to the “AlexNet” \([24]\).
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.

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,aft er,which,evidence,was,pulmonary,mead,adrenal, prior,pelvic,without,cysts,spleen,mass,disease,multiple,isovue-300,obtained,areas,consisten,t nodules,changes,pleural,lesions,following,abdominal,that,hilar,axillary].
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 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.

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

<table>
<thead>
<tr>
<th></th>
<th>AlexNet 8-layers</th>
<th>VGG 19-layers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV</td>
<td>top-1</td>
</tr>
<tr>
<td>document-level</td>
<td>0.6078</td>
<td>0.0692</td>
</tr>
<tr>
<td>document-level-h2</td>
<td>0.3448</td>
<td>0.3252</td>
</tr>
<tr>
<td>sentence-level</td>
<td>0.48</td>
<td>0.48</td>
</tr>
</tbody>
</table>

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-weighted⁶], [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.

⁶Natural language expressions for imaging modalities of magnetic resonance imaging (MRI).
<table>
<thead>
<tr>
<th>#words/sentence</th>
<th>mean</th>
<th>median</th>
<th>std</th>
<th>max</th>
<th>min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report-wide</td>
<td>17.7</td>
<td>9</td>
<td>8.99</td>
<td>104</td>
<td>1</td>
</tr>
<tr>
<td>Image references</td>
<td>23.22</td>
<td>19</td>
<td>16.99</td>
<td>221</td>
<td>4</td>
</tr>
<tr>
<td>Image references, no stopwords no digits</td>
<td>13.46</td>
<td>11</td>
<td>9.94</td>
<td>143</td>
<td>2</td>
</tr>
<tr>
<td>Image references, disease terms only</td>
<td>3.17</td>
<td>4</td>
<td>2.52</td>
<td>25</td>
<td>1</td>
</tr>
</tbody>
</table>

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 workflow 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 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) 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]:

$$E = -\frac{1}{n} \sum_{n=1}^{N} [g(z_n)g(\tilde{z}_n) + (1 - g(z_n)) \log(1 - g(\tilde{z}_n))],$$

(2)

where $z_n$ or $\tilde{z}_n$ is any uni-element of the target word vectors $Z_n$ or optimized output vectors $\tilde{Z}_n$, $g(x)$ is the sigmoid function.

---

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 \( g(x) = \frac{1}{1 + e^x} \), 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 \( 256 \times 2 \) (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 \( \mathbb{R}^{256 \times 2} \) 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 \( \mathbb{R}^{256 \times 2} \) 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.

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


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


