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

Various patient data of a large population are available on the patient archive and communication system (PACS) of many hospitals or clinical institutions. However such data are not widely studied, due to the challenges encountered in analyzing a large clinical dataset. Nonetheless, efficient analysis of large data can lead us to gain useful, possibly unpreceded insights in the area under study. With the big-data analysis on large collection of radiology images, we aim to achieve 'predictive medicine' – detecting diseases with large population patient image screening.

In this paper, we show a first attempt to achieve this. We collected about 780,000 radiology reports from the PACS of the National Institutes of Health Clnical Center, comprising about 1 billion words. However, manually examining and annotating these is not only challenging, but also requires an expertise in radiology. To fill in this gap, we used a non-parametric topic modeling algorithm (Latent Dirichlet Allocation (LDA) [1]), to analyze the large collection of reports and to divide them into a number of categories with semantic hierarchies (document-level topics, document-level-2nd-hierarchy topics, and sentence-level topics). Each category is defined based on the unique occurrences of "key words" contributing to the category, and with natural language processing (NLP) we could find the images mentioned in the reports.

We found that the categories modeled by LDA largely correlate with the images, e.g. a "brain tumor" category has mostly MR images of brain, and a "breast imaging" category has mostly breast mammograms. Some examples of image-topic associations and key-words are shown in Figure 1.

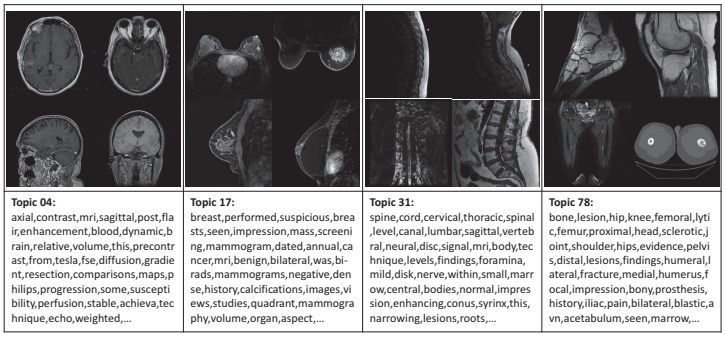

Figure 1: Examples of LDA generated document-level topics with corresponding images and key words. Topic #4 MRI of brain tumor; #17 breast imaging; #31 degenerative spine disc disease; #78 bone metastases.

We then trained convolutional neural networks (CNNs) to classify the images into the report categories, e.g. to classify whether an image is about "breast imaging", or "bone metastasis". Two popular CNN basis models were used ("AlexNet" [4] and "VGG-19" [7]), where we fine-tuned the parameters trained on the ImageNet [2] dataset using Caffe [3] framework. While the images of some topic categories and some body parts are easily distinguishable, the visual differences in abdominal parts are rather subtle. 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, and the classification accuracies in Table 1 support this.

	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 1: Validation and top-1, top-5 test scores in classification accuracy using AlexNet [4] and VGG-19 [7] deep CNN models.

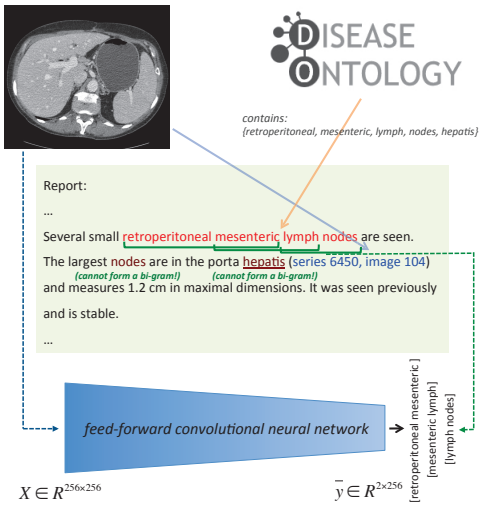

Figure 2: An example illustration of how word sequences are learned for an image. Bi-grams are selected from the image's reference sentences containing disease-related terms from the disease ontology (http://disease-ontology.org), and each bi-gram is converted to a vector of $\mathbf{Z} \in \mathbb{R}^{256\times 2}$ to learn from an image.

We then train a new set of CNNs to predict the "key words" mentioning the images, e.g. to predict "adenopathy", "masses", "lung", given a CT image with lung cancer. The steps involved for generating text for a radiology image are: *(1)* map words to vectors using word2vec model [5, 6]; *(2)* mine and match image to description relation, by extracting disease-related bi-grams in the sentences of report mentioning images; *(3)* convert the bi-grams to vectors, and train CNNs to map images to the word-vectors describing the image; *(4)* for the (three-level) topics the image is classified to, find the key words for each topic which are closest to the output vectors using cosine distance. A brief illustration of how we train the image-to-description models is shown in Figure 1. 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.

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 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 [2] for computer vision.

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

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *http://caffe.berkeleyvision.org*.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.