EXPLOITING VISUAL AND REPORT-BASED INFORMATION FOR CHEST X-RAY ANALYSIS BY JOINTLY LEARNING VISUAL CLASSIFIERS AND TOPIC MODELS

Zachary A. Daniels and Dimitris N. Metaxas

Department of Computer Science, Rutgers University, Piscataway, NJ, USA

ABSTRACT

Manual examination of chest x-rays is a time consuming process that involves significant effort by expert radiologists. Recent work attempts to alleviate this problem by developing learning-based automated chest x-ray analysis systems that map images to multi-label diagnoses using deep neural networks. These methods are often treated as black boxes, or they output attention maps but don't explain why the attended areas are important. Given data consisting of a frontal-view x-ray, a set of natural language findings, and one or more diagnostic impressions, we propose a deep neural network model that during training simultaneously 1) constructs a topic model which clusters key terms from the findings into meaningful groups, 2) predicts the presence of each topic for a given input image based on learned visual features, and 3) uses an image's predicted topic encoding as features to predict one or more diagnoses. Since the net learns the topic model jointly with the classifier, it gives us a powerful tool for understanding which semantic concepts the net might be exploiting when making diagnoses, and since we constrain the net to predict topics based on expert-annotated reports, the net automatically encodes some higher-level expert knowledge about how to make diagnoses.

***Index Terms*—** Chest X-Ray Analysis, Multimedia Analysis, Natural Language Processing, Deep Learning

1 INTRODUCTION

Visual inspection of chest x-rays is a common and important method for diagnosing certain life-threatening diseases such as pneumonia, but manual examination of chest x-rays is time-consuming, requiring significant effort by highly-trained radiologists. (Semi-)automated chest x-ray analysis using computer vision and machine learning algorithms can act as a support tool for radiologists, allowing them to make faster diagnoses and spend more time focusing on difficult cases. Recent work in automated chest x-ray analysis focuses on learning to map from images to multi-label diagnoses using deep neural networks (DNNs) (e.g. [1, 2, 3]). These models are often treated as black boxes, or they output attention maps that show where the net is focusing when predicting a diagnosis but do not explain why these areas are important. The nets are also not designed to incorporate expert knowledge, so they must learn to make diagnoses "from scratch".

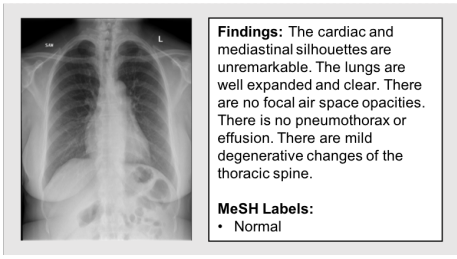

Fig. 1. An example ground-truth report

Many chest x-rays come with radiologist-dictated reports that justify the diagnostic impressions. Natural language processing (NLP) has been used extensively to analyze biomedical text (e.g. [4, 5, 6, 7]) but until recently, there has been little focus on integrating rich information from reports with powerful DNNs to improve automated biomedical image analysis (e.g. [8, 9, 10]) and in some cases, produce reports directly from images (e.g. [11, 12, 13, 14]). In this work, we propose a new method that exploits both visual and textual information for improved automated chest x-ray analysis.

1.1 Proposed Approach and Contributions

Given frontal-view x-rays, a set of corresponding natural language findings, and one or more corresponding MeSH labels (Medical Subject Headings, annotations using a controlled vocabulary) (example in Fig. 1), we propose a DNN architecture that *during training* simultaneously 1) **constructs a topic model** (see [15]) which clusters key terms from the findings into meaningful groups (e.g. "lungs", "clear", and "expanded" might form a topic), 2) **predicts the presence or absence of each topic** for a given image based on learned visual features, and 3) uses an image's predicted topic encoding as features for **predicting one or more diagnoses**. *At test time*, only images are needed as input. Since the DNN learns the topic model jointly with the classifier, it gives us a powerful tool for investigating which semantic concepts the net might be exploiting when making diagnoses. Since the net is constrained to predict topics based on expert-annotated

This work is partly supported by the Air Force Office of Scientific Research (AFOSR) under the Dynamic Data-Driven Application Systems Program, NSF 1763523, 1747778, 1733843 and 1703883, and NIH 1R01HL127661-01 Awards.

reports and then use these topics to predict diagnoses, we force the net to "think" like an expert, encouraging it to learn higher-level features that it might have otherwise missed.

In the next section, we discuss 1) how to extract an initial set of key terms from the natural language reports, 2) how to learn a dictionary of topics and an encoding vector for each report based on Pseudo-Boolean Matrix Factorization, and 3) how to integrate topic modeling into DNN architectures.

2 METHODOLOGY

2.1 Pre-Training the Feature Extraction Neural Network

Our experiments are conducted on frontal-view chest x-ray images from the OpenI dataset. Training DNNs require large amounts of data, and the subset of OpenI used only includes 3,821 images, so we first pre-train a ResNet-22 model [16] on the larger ChestX-Ray14 dataset [1] and subsequently finetune the network on the OpenI dataset. To train the initial network, we use a variant of the multi-label cross entropy as our loss function:

$$\mathcal{L}_c = \sum_{l \in L} \left(\frac{1}{|P_l|} \sum_{y_{li} \in P_l} -y_{li} \log \hat{y}_{li} + \frac{1}{|N_l|} \sum_{y_{lj} \in N_l} -(1 - y_{lj}) \log (1 - \hat{y}_{lj}) \right)$$

L is the set of all labels. l is an individual label. P_l and N_l are the sets of positive examples and negative examples in a batch for label l, respectively. $|X|$ denotes the number of examples in a batch for set X. $y_{li} \in \{0, 1\}$ is the label of ith positive instance of label l, and $y_{lj} \in \{0, 1\}$ represents the jth negative instance. $\hat{y}_{li} \in [0, 1]$ and $\hat{y}_{lj} \in [0, 1]$ are the predicted scores of instances i and j.

The data is often imbalanced where a label will have many more negative than positive examples. To address this problem, 1) we weigh the entropies for the positive and negative examples by $\frac{1}{|P_l|}$ and $\frac{1}{|N_l|}$, and 2) we construct each mini-batch using *stratified sampling with replacement*. For each mini-batch, we select one label, and then randomly select positive instances for this label for half of the batch and negative examples for the other half. After every batch, we select another label, and iterate through all of the labels. This procedure allows the net to see instances with rare labels more frequently than when standard sampling methods are used.

We use a batch size of 32 with 2,703 total minibatches per epoch. We train for 50 epochs. We also augment the data using random cropping and by making small random adjustments to the brightness, contrast, and saturation of the cropped images. We use images of size 512-by-512 pixels.

2.2 Extracting Key Terms from Natural Language Text

The goal of this project is to use information captured in natural language text reports to help train DNN-based models to find more meaningful visual features. To achieve this goal, we propose a DNN that learns to predict a set of labels representing diagnoses while simultaneously constructing a topic model consisting of 1) a dictionary which clusters related key terms together and 2) encoding vectors that capture the presence or absence of each topic in each input instance. In this section, we focus on how to extract an initial set of key terms from a database of reports.

For each document, simple *rule-based negative scope detection* is applied to capture negation, so phrases like "no pleural effusion" are parsed as "pleural effusion_neg". Next, stop words (i.e. common words like "the") are removed. Then, SGRank [17] is used to identify important n-grams (e.g. "pleural effusion", "focal airspace disease", "pneumothorax", etc.). Finally, we extract a *bag-of-key terms (BOKT)* representation from all documents. After pruning terms that appear fewer than ten times, 600-700 terms remain per fold.

2.3 Learning Topic Models using Matrix Factorization

The naïve way to incorporate text information into a visual DNN is to directly predict the BOKT for each image. Several problems exist with this approach. 1) The key term extraction process is not perfect, so "pleural effusion", "pleural effusions", and "effusion" might be extracted as different terms, making it difficult to learn a classifier for each individual term. 2) Synonyms and abbreviations present similar problems, e.g. "copd" and "chronic obstructive pulmonary disease". 3) Sometimes individual terms are not useful by themselves but become useful when paired, e.g. "cardiac silhouette" becomes useful when paired with "unremarkable". 4) Terms can be redundant due to co-occurrence, e.g. "normal cardiac silhouette" and "normal mediastinum size" frequently appear together. 5) There are limited training examples for each term, making it difficult to learn fine-grained classifiers. Instead of considering individual key terms, we can exploit context between terms to form a lower-dimensional set of topics, and use the DNN to predict these topics.

Suppose we have a binary matrix A where each row represents a key term and each column represents a document (report). A_{ij} is 1 if key term i is present in document j and 0, otherwise. We can factorize this matrix into a dictionary matrix D which clusters related key terms into groups and an encoding matrix E which tells us which topics are present in each document. This problem can be modeled by Boolean Matrix Factorization (BMF) [18] which assumes the document-term, dictionary, and encoding matrices are binary and Boolean matrix multiplication is satisfied. BMF forces hard assignments of terms to topics which promotes easier interpretation of the topics, and documents are formed by taking the union of the terms of its constituent topics. BMF is a computationally challenging, so we utilize a relaxation, **Pseudo-Boolean Matrix Factorization (PBMF)** [19]:

$$\min_{D,E} ||\Omega \bullet (A - \min(DE, 1 + 0.01DE))||_2 + \alpha_1 ||E||_1 + \alpha_2 ||D||_1 + \alpha_3 ||D^T D - diag(D^T D)||_2 \; s.t. \; D, E \in [0, 1] \quad (1)$$

The first term reconstructs the document-term matrix using

approximate Boolean matrix multiplication. The second and third terms encourage sparsity in the encoding and dictionary matrices. By promoting orthogonality between dictionary vectors, the final term forces similar topics to merge, leading to a more concise representation. Generally, sparsity and orthogonality help to improve the interpretability of the resulting topic model, but this comes at a cost. As the model becomes more constrained (i.e., we increase the values of the αs), it also becomes less expressive, which can hurt performance on the target task. Finally, we try to force the model to pay attention to rare key terms by adjusting the reconstruction error using the inverse document frequency (idf): $\Omega = max(idf(A) \bullet A, 0.25), idf_{term} = (1 + log\left(\frac{N_{all}}{N_{term}}\right))$ where N_{all} is the number of instances and N_{term} is the number of instances containing a specific term.

2.4 Incorporating Topic Modeling into Convolutional Neural Networks

We can reformulate the PBMF optimization problem (Eq. 1) as a loss function $\mathcal{L}_t$. The encoding $E = sigmoid(f_E(X, W_E))$ becomes a function of the input training images X and learnable network parameters W_E, and the dictionary is a constrained variable of the network $D = sigmoid(W_D)$. Combining the topic model loss (Eq. 1) with the classification loss, we get a joint loss: $\mathcal{L} = \mathcal{L}_t + \beta\mathcal{L}_c$.

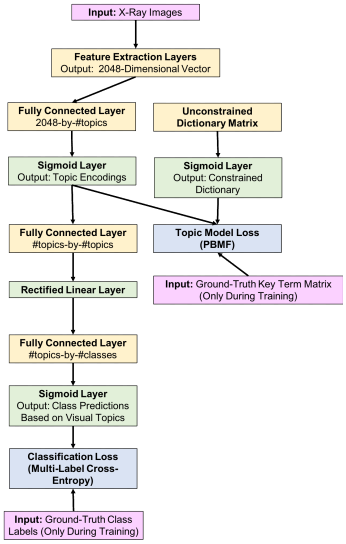

Fig. 2. Structure of the proposed neural network model

We make minor changes to the architecture from [19] which uses a DNN to predict the topic encoding for each training instance, updates the dictionary in response to the predicted topic encodings, and uses the topic encodings as features for some target classification task. By bottlenecking through the topic modeling layer, the net is forced to "think" like an expert. Unlike [19], we replace the linear classification block with a non-linear block. We also simultaneously optimize W_D and W_E instead of performing alternating optimizing. We still solve for an initial D outside of the network and finetune D inside the net, leading to faster convergence to a better topic dictionary. Our architecture appears in Fig. 2.

We train a net for 100 epochs with 166 batches per epoch using a batch size of 16 using the sampling and data augmentation strategies as discussed in Section 2.1. We use $\alpha_1 = 0.01, \alpha_2 = 0.001, \alpha_3 = 0.1$, and $\beta = 10$ which were determined experimentally on the first fold using a holdout set. In future work, we will consider better methods for tuning these parameters (specifically Bayesian optimization).

3 EXPERIMENTS

3.1 Datasets and Experimental Procedure

For learning the pre-trained net, we use the 86,524 training images from the ChestX-Ray14 dataset [1] using all 14 labels. For training and evaluating the final model, we use 3,821 frontal-view images from the OpenI dataset [20] along with their corresponding "FINDINGS" annotations. We align the labels in the OpenI dataset with those of the ChestX-Ray14 dataset, and keep the 5 shared labels that appear most often: "normal/no findings", "atelecstasis", "cardiomegaly", "effusion", and "emphysema". We use 5-fold cross validation with splits based on individual patients (not individual images).

3.2 Evaluation Metrics

The most common metric for evaluating automated Chest X-Ray analysis systems is the *macro-area under the receiver-operator curve (macro-AUROC)*, but this metric can be misleading because common negative labels are treated with equal importance to rare positive labels. In practice, identifying positive cases is generally a higher priority, so metrics based on the precision-recall curve should be used instead. The *mean average precision (mAP)* summarizes the PR-curve over all labels. Sometimes it is useful to rank the potential labels in order of most to least likely for further inspection by a human expert. To evaluate such rankings, we consider the *multi-label ranking average precision (mlrAP)* which summarizes how often true labels are ranked higher than false labels and the *multi-label coverage* which summarizes on average how many labels need to be inspected for a given instance before all of the true labels have been inspected.

Features	m-AUROC	mAP	Coverage	mlrAP
All Key Terms	0.969	0.814	0.662	0.980
Doc2Vec	0.927	0.613	0.718	0.962
Our Approach	0.928	0.674	0.710	0.963

Table 1. Results based on training a classifier on features extracted from the natural language reports

3.3 Text-Based Experiments

To see how much information is captured in the reports without considering image data, we train classifiers using

text-based features. We report the results in Table 1. Using the raw BOKT representation as our features, we achieve high macro-AUROC (~0.97) and mAP(~0.81) and coverage and mlrAP are close to perfect. As a baseline NLP-based dimensionality reduction method, we use the distributed bag-of-words Doc2Vec embedding [21] with 200 dimensions trained on the reports. This is a common document-level embedding method. Finally, we evaluate our approach: a PBMF topic model with 200 topics. Both our approach and the Doc2Vec representation lose a significant amount of information compared to the raw BOKT, but results are still promising. Our representation matches Doc2Vec in terms of macro-AUROC, coverage, and mlrAP, but outperforms it by ~6% in terms of mAP.

3.4 Imaging-Based Experiments

Next, we consider how well we can predict a set of diagnostic labels using visual inputs. We train and test a standard ResNet-22 and a modified ResNet-22 that bottlenecks through the topic modeling layer (using 200 topics). We report the global statistics and statistics for the individual labels in Table 2. Our approach globally outperforms the standard model by about 1-2% in both the macro-AUROC and mAP, and outperforms the standard model in both AUROC and AP for three of the five diagnostic labels. This a relatively small improvement compared to what Table 1 suggests should be possible. This is because the topic recognition component is imperfect. If we compare our visually recognized topic encodings to *approximate* ground-truth encodings (based on the dictionary learned by the net and thresholded at 0.5), we only achieve an average mAP of ~ 0.24 when considering all 200 topics. The net overfits w.r.t. rare topics. If we only consider topics that appear in at least 75 ($\sim 2\%$) of the training instances which accounts for $\sim 60\%$ of the learned topics, the mAP for topic recognition significantly rises to ~ 0.39.

	Standard	Our Approach
macro-AUROC	0.857	0.867
mAP	0.459	0.477
Coverage	0.825	0.851
mlrAP	0.927	0.925

		Standard		Our Approach	
Diagnosis	# Instances	AUROC	AP	AUROC	AP
Normal	1395	0.774	0.647	0.783	0.655
Atelectasis	309	0.785	0.303	0.812	0.332
Cardiomegaly	329	0.930	0.592	0.927	0.587
Effusion	148	0.926	0.514	0.921	0.535
Emphysema	110	0.868	0.241	0.892	0.276

Table 2. Performance of learned visual classifiers: 1) overall (top) and 2) on individual diagnostic labels (bottom). We compare a standard ResNet-22 architecture with our modified architecture that bottlenecks through the topic modeling layer.

3.5 Analysis of Quantitative Results

The proposed model's biggest weakness is its ability to overfit to rare (and noisy) topics, leading to suboptimal performance on the test data. This weakness can be overcome in several ways, e.g., by 1) less noisy extraction of key terms; 2) pruning visually meaningless topics; 3) collecting annotations that are more information-complete; and/or 4) collecting more data. We will consider such improvements in future work. Despite the imperfect topic recovery from visual data, we still see some minor improvement in performance on the target task. There are several possible explanations for this. First, our approach utilizes additional privileged information during training in the form of expert-annotated "findings". This potentially forces the net to focus on different types of discriminative features that are not obvious from visual inspection alone. Second, having to jointly learn the topic model and classifier might act as a form of regularization, preventing the network from overfitting. Lastly, there are minor differences in the base network architecture (e.g. our architecture utilizes a non-linear classification block) which might affect performance.

3.6 Qualitative Results

Fig. 3 shows an example of an x-ray and its highly-ranked topics, demonstrating some of the types of concepts the net attempts to recognize and showing the utility and potential of using topics as an interpretable intermediate feature layer. Bottlenecking through the topic modeling layer is useful in 1) helping the net discover more discriminative features and 2) understanding what the net is attempting to learn. With more data, the topic encodings should be predicted with greater accuracy and robustness, leading to improved interpretability.

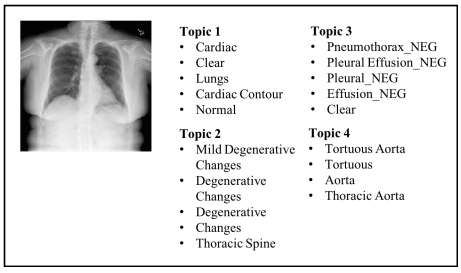

Fig. 3. Example of an x-ray and its highly-ranked topics

4 CONCLUSIONS AND FUTURE WORK

We proposed a model that utilized report-based and visual information in a deep learning framework for automated chest x-ray analysis. We validated the utility of the approach experimentally, and discussed how it can be useful for developing systems which incorporate expert knowledge and are easier to analyze. In future work, we intend to address limitations relating to the size of the training data, imperfect NLP, and uncertainty and incompleteness of the annotations, while also exploring how multi-view data can be utilized, how patient history and demographics can be incorporated, and how we can generate sentences from the topic encoding.

5 REFERENCES

[1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*. IEEE, 2017, pp. 3462–3471.

[2] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[3] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[4] Peter Spyns, "Natural language processing in medicine: an overview," *Methods of information in medicine*, vol. 35, no. 04/05, pp. 285–301, 1996.

[5] Martin Krallinger and Alfonso Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome biology*, vol. 6, no. 7, pp. 224, 2005.

[6] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors, "Natural language processing in radiology: a systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[7] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al., "Clinical information extraction applications: a literature review," *Journal of biomedical informatics*, 2017.

[8] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers, "Interleaved text/image deep mining on a very large-scale radiology database," in *CVPR*, 2015, pp. 1090–1099.

[9] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers, "Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3729–3759, 2016.

[10] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *CVPR*. IEEE, 2018, pp. 9049–9058.

[11] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *CVPR*, 2016, pp. 2497–2506.

[12] Baoyu Jing, Pengtao Xie, and Eric Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[13] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *CVPR*, 2017, pp. 6428–6436.

[14] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *MICCAI*. Springer, 2018, pp. 457–466.

[15] David M Blei and John D Lafferty, "Topic models," in *Text Mining*, pp. 101–124. Chapman and Hall/CRC, 2009.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] Soheil Danesh, Tamara Sumner, and James H Martin, "Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction," in **SEM*, 2015, pp. 117–126.

[18] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila, "The discrete basis problem," *IEEE TKDE*, 2008.

[19] Zachary A Daniels and Dimitris Metaxas, "Scenarionet: An interpretable data-driven model for scene understanding," *IJCAI Workshop on XAI*, p. 33, 2018.

[20] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2015.

[21] Quoc Le and Tomas Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014, pp. 1188–1196.