TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays

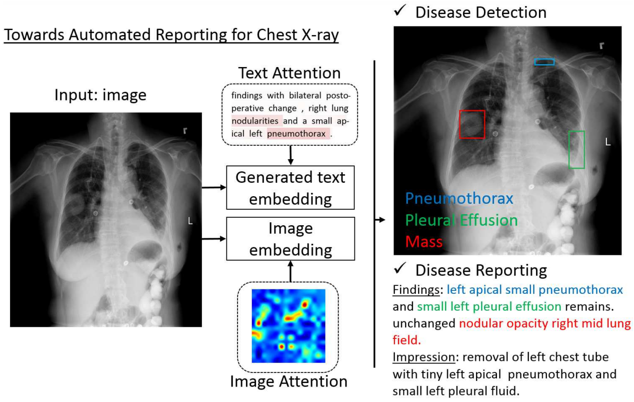
Xiaosong Wang*[1], Yifan Peng*[2], Le Lu[1], Zhiyong Lu[2], Ronald M. Summers[1]
[1]Department of Radiology and Imaging Sciences, Clinical Center,
[2] National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20892
{xiaosong.wang,yifan.peng,luzh,rms}@nih.gov, lel@nvidia.com

Abstract

Chest X-rays are one of the most common radiological examinations in daily clinical routines. Reporting thorax diseases using chest X-rays is often an entry-level task for radiologist trainees. Yet, reading a chest X-ray image remains a challenging job for learning-oriented machine intelligence, due to (1) shortage of large-scale machine-learnable medical image datasets, and (2) lack of techniques that can mimic the high-level reasoning of human radiologists that requires years of knowledge accumulation and professional training. In this paper, we show the clinical free-text radiological reportscan be utilized as a priori knowledge for tackling these two key problems. We propose a novel Text-Image Embedding network (TieNet) for extracting the distinctive image and text representations. Multi-level attention models are integrated into an end-to-end trainable CNN-RNN architecture for highlighting the meaningful text words and image regions. We first apply TieNet to classify the chest X-rays by using both image features and text embeddings extracted from associated reports. The proposed auto-annotation framework achieves high accuracy (over 0.9 on average in AUCs) in assigning disease labels for our hand-label evaluation dataset. Furthermore, we transform the TieNet into a chest X-ray reporting system. It simulates the reporting process and can output disease classification and a preliminary report together. The classification results are significantly improved (6% increase on average in AUCs) compared to the state-of-the-art baseline on an unseen and hand-labeled dataset (OpenI).

1. Introduction

In the last decade, challenging tasks in computer vision have gone through different stages, from sole image classification to multi-category multi-instance classification/detection/segmentation to more complex cognitive tasks that involve understanding and describing the relationships of object instances inside the images or videos. The rapid and significant performance improvement is partly driven by publicly accessible of the large-scale image and video datasets with quality annotations, *e.g.*, ImageNet [8], PASCAL VOC [10], MS COCO [22], and Visual Genome [18] datasets. In particular, ImageNet pretrained deep Convolutional Neural Network (CNN) models [15, 19, 21] has become an essential basis (indeed an advantage) for many higher level tasks, *e.g.*, Recurrent Neural Network (RNN) based image captioning [34, 17, 30, 11], Visual Question Answering [36, 42, 38, 27], and instance relationship extraction [16, 14, 6].



findings with bilateral postoperative change , right lung nodularities and a small apical left pneumothorax .

Findings: left apical small pneumothorax and small left pleural effusion remains. unchanged nodular opacity right mid lung field.

Impression: removal of left chest tube with tiny left apical pneumothorax and small left pleural fluid.

Figure 1. Overview of the proposed automated chest X-ray reporting framework. A multi-level attention model is introduced.

On the contrary, there are few publicly available large-scale image datasets in the medical image domain. Conventional means of annotating natural images, e.g crowd-sourcing, cannot be applied to medical images due to the fact that these tasks often require years of professional training and domain knowledge. On the other hand, radiological raw data (*e.g.*, images, clinical annotations, and radio-

*Both authors contributed equally.

logical reports) have been accumulated in many hospitals' Picture Archiving and Communication Systems (PACS) for decades. The main challenge is how to transform those retrospective radiological data into a machine-learnable format. Accomplishing this with chest X-rays represents a major milestone in the medical-imaging community [35].

Different from current deep learning models, radiologists routinely observe multiple findings when they read medical images and compile radiological reports. One main reason is that these findings are often correlated. For instance, liver metastases can spread to regional lymph nodes or other body parts. By obtaining and maintaining a holistic picture of relevant clinical findings, a radiologist will be able to make a more accurate diagnosis. To our best knowledge, developing a universal or multi-purpose CAD framework, which is capable of detecting multiple disease types in a seamless fashion, is still a challenging task. However, such a framework is a crucial part to build an automatic radiological diagnosis and reporting system.

Toward this end, we investigate how free-text radiological reports can be exploited as *a priori* knowledge using an innovative text-image embedding network. We apply this novel system in two different scenarios. We first introduce a new framework for auto-annotation of the chest X-rays by using both images features and text embeddings extracted from associated reports. Multi-level attention models are integrated into an end-to-end trainable CNN-RNN architecture for highlighting the meaningful text words and image regions. In addition, we convert the proposed annotation framework into a chest X-ray reporting system (as shown in Figure 1). The system stimulates the real-world reporting process by outputting disease classification and generating a preliminary report spontaneously. The text embedding learned from the retrospective reports are integrated into the model as *a priori* knowledge and the joint learning framework boosts the performance in both tasks in comparison to previous state-of-the-art.

Our contributions are in fourfold: (1) We proposed the Text-Image Embedding Network, which is a multi-purpose end-to-end trainable multi-task CNN-RNN framework; (2) We show how raw report data, together with paired image, can be utilized to produce meaningful attention-based image and text representations using the proposed TieNet. (3) We outline how the developed text and image embeddings are able to boost the auto-annotation framework and achieve extremely high accuracy for chest x-ray labeling; (4) Finally, we present a novel image classification framework which takes images as the sole input, but uses the paired text-image representations from training as *a prior* knowledge injection, in order to produce improved classification scores and preliminary report generations.

Importantly, we validate our approach on three different datasets and the TieNet improves the image classification result (6% increase on average in area under the curve (AUC) for all disease categories) in comparison to the state-of-the-art on an unseen and hand-labeled dataset (OpenI [7]) from other institute. Our multi-task training scheme can help not only the image classification but also the report generation by producing reports with higher BLEU scores than the baseline method.

2. Related work

Computer-Aided Detection (CADe) and Diagnosis (CADx) has long been a major research focus in medical image processing [5]. In recent years, deep learning models start to outperform conventional statistical learning approaches in various tasks, such as automated classification of skin lesions [9], detection of liver lesions [4], and detection of pathological-image findings [40]. However, current CADe methods typically target one particular type of disease or lesion, such as lung nodules, colon polyps or lymph nodes [24].

Wang *et al*. [35] provide a recent and prominent exception, where they introduced a large scale chest X-ray dataset by processing images and their paired radiological reports (extracted from their institutional PACS database) with natural language processing (NLP) techniques. The publicly available dataset contains $112,120$ front-view chest X-ray images of $30,805$ unique patients.However, radiological reports contain richer information than simple disease binary labels, *e.g*., disease location and severity, which should be exploited in order to fully leverage existing PACS datasets. Thus, we differ from Wang *et al*.'s approach by leveraging this rich text information in order to produce an enhanced system for chest X-ray CADx.

In vision of visual captioning, our work is closed to [37, 33, 29, 38, 27]. Xu *et al*. [37] first introduced the sequence-to-sequence model and spatial attention model into the image captioning task. They conditioned the long short-term memory (LSTM) decoder on different parts of the input image during each decoding step, and the attention signal was determined by the previous hidden state and CNN features. Vinyals *et al*. [33] cast the syntactical parsing problem as a sequence-to-sequence learning task by linearizing the parsing tree. Pederoli *et al*. [29] allowed a direct association between caption words and image regions. More recently, multi-attention models [38, 27] extract salient regions and words from both image and text and then combine them together for better representations of the pair. In medical imaging domain, Shin *et al*.[32] proposed to correlate the entire image or saliency regions with MeSH terms. Promising results [41] are also reported in summarizing the findings in pathology images using task-oriented reports in the training. The difference between our model and theirs lies in that we employ multi-attention models with a mixture of image and text features in order to provide more

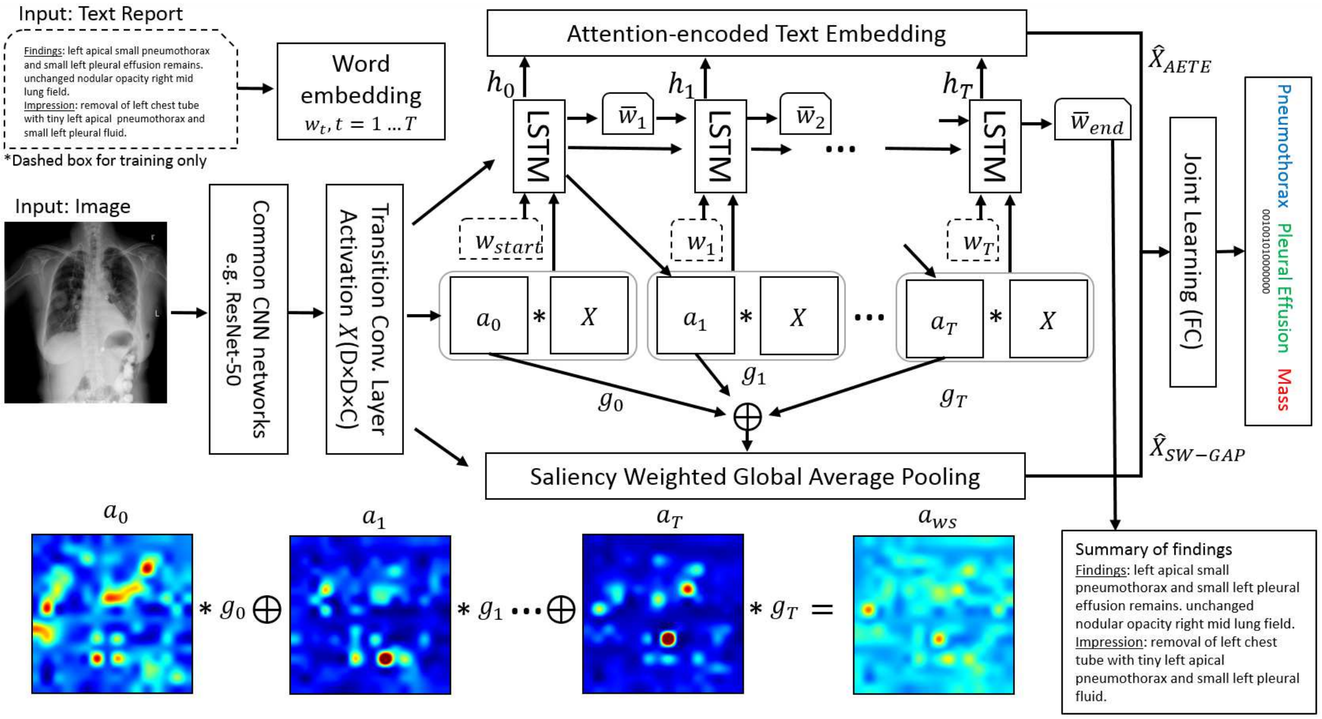

Input: Text Report
Findings: left apical small pneumothorax and small left pleural effusion remains. unchanged nodular opacity right mid lung field.
Impression: removal of left chest tube with tiny left apical pneumothorax and small left pleural fluid.
*Dashed box for training only
Word embedding $w_t, t = 1 \ldots T$
Input: Image
Common CNN networks e.g. ResNet-50
Transition Conv. Layer Activation X(D×D×C)
Attention-encoded Text Embedding
h_0
h_1
h_T
LSTM
$\bar{w}_1$
$\bar{w}_2$
$\bar{w}_{end}$
w_{start}
w_1
w_T
a_0 * X
a_1 * X
a_T * X
g_0
g_1
g_T
Saliency Weighted Global Average Pooling
$\hat{X}_{AETE}$
$\hat{X}_{SW-GAP}$
Joint Learning (FC)
Pneumothorax Pleural Effusion Mass
0010101000000
a_0
a_1
a_T
a_{ws}
$* g_0 \oplus$
$* g_1 \cdots \oplus$
$* g_T =$
Summary of findings
Findings: left apical small pneumothorax and small left pleural effusion remains. unchanged nodular opacity right mid lung field.
Impression: removal of left chest tube with tiny left apical pneumothorax and small left pleural fluid.

Figure 2. Framework of the proposed chest X-ray auto-annotation and reporting framework. Multi-level attentions are introduced to produce saliency-encoded text and image embeddings.

salient and meaningful embeddings for the image classification and report generation task.

Apart from visual attention, text-based attention has also been increasingly applied in deep learning for NLP [2, 26, 31]. It attempts to relieve one potential problem that the traditional encoder-decoder framework faces, which is that the input is long or very information-rich and selective encoding is not possible. The attention mechanism attempts to ease the above problems by allowing the decoder to refer back to the input sequence [39, 23, 25]. To this end, our work closely follows the one used in [23] where they extracted an interpretable sentence embedding by introducing self-attention. Our model paired both the attention-based image and text representation from training as a prior knowledge injection to produce improved classification scores.

3. Text-Image Embedding Network

The radiological report is a summary of all the clinical findings and impressions determined during examination of a radiography study. A sample report is shown in Figure 1. It usually contains richer information than just disease keywords, but also may consist of negation and uncertainty statements. In the 'findings' section, a list of normal and abnormal observations will be listed for each part of the body examined in the image. Attributes of the disease patterns, *e.g.*, specific location and severity, will also be noted. Furthermore, critical diagnosis information is often presented in the 'impression' section by considering all findings, patient history, and previous studies. Suspicious findings may cause recommendations for additional or follow-up imaging studies. As such, reports consist of a challenging mixture of information and a key for machine learning is extracting useful parts for particular applications.

In addition to mining the disease keywords [35] as a summarization of the radiological reports, we want to learn a text embedding to capture the richer information contained in raw reports. Figure 2 illustrates the proposed Text-Image Embedding Network. We first introduce the foundation of TieNet, which is an end-to-end trainable CNN-RNN architecture. Afterwards we discuss two enhancements we develop and integrate, *i.e.*, attention-encoded text embedding (AETE) and saliency weighted global average pooling (SW-GAP). Finally, we outline the joint learning loss function used to optimize the framework.

3.1. End-to-End Trainable CNN-RNN Model

As shown in Figure 2, our end-to-end trainable CNN-RNN model takes an image I and a sequence of 1-of-V encoded words.

$$\mathbf{S} = \{\mathbf{w}_1, \ldots, \mathbf{w}_T\}, \mathbf{w}_t \in \mathbb{R}^V, \quad (1)$$

where $\mathbf{w}_t$ is a vector standing for a d_w dimensional word embedding for the t-th word in the report, V is the size of the vocabulary, and T is the length of the report. The initial CNN component uses layers borrowed from ImageNet pre-trained models for image classification, *e.g.*, ResNet-50 (from Conv1 to Res5c). The CNN component additionally includes a convolutional layer (transition layer) to manipulate the spatial grid size and feature dimension.

Our RNN is based off of Xu *et al.*'s visual image spatial attention model [37] for image captioning. The convolutional activations from the transition layer, denoted as $\mathbf{X}$, initialize the RNN's hidden state, $\mathbf{h}_t$, where a fully-connected embedding, $\phi(\mathbf{X})$, maps the size d_X transition layer activations to the LSTM state space of dimension d_h. In addition, X is also used as one of the RNN's input. However, following Xu *et al.* [37], our sequence-to-sequence model includes a deterministic and soft visual spatial attention, $\mathbf{a}_t$, that is multiplied element-wise to $\mathbf{X}$ before the latter is inputted to the RNN. At each time step, the RNN also outputs the subsequent attention map, $\mathbf{a}_{t+1}$.

In addition to the soft-weighted visual features, the RNN also accepts the current word at each time step as input. We adopt standard LSTM units [13] for the RNN. The transition to the next hidden state can then be denoted as

$$\mathbf{h}_t = LSTM([\mathbf{w}_t, \mathbf{a}_t, \mathbf{X}], \mathbf{h}_{t-1}). \tag{2}$$

The LSTM produces the report by generating one word at each time step conditioned on a context vector, *i.e.*, the previous hidden state $\mathbf{h}_t$, the previously generated words $\mathbf{w}_t$, and the convolutional features of $\mathbf{X}$ whose dimension is $D \times D \times C$. Here $D = 16$ and $C = 1024$ denote the spatial and channel dimensions, respectively. Once the model is trained, reports for a new image can be generated by sequentially sampling $\mathbf{w}_t \sim p(\mathbf{w}_t|\mathbf{h}_t)$ and updating the state using Equation 2.

The end-to-end trainable CNN-RNN model provides a powerful means to process both text and images. However, our goal is also to obtain an interpretable global text and visual embedding for the purposes of classification. For this reason, we introduce two key enhancements in the form of the AETE and SW-GAP.

3.2. Attention Encoded Text Embedding

To compute a global text representation, we use an approach that closely follows the one used in [23]. More specifically, we use attention to combine the most salient portions of the RNN hidden states. Let $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_T)$ be the $d_h \times T$ matrix of all the hidden states. The attention mechanism outputs a $r \times T$ matrix of weights $\mathbf{G}$ as

$$\mathbf{G} = softmax(\mathbf{W}_{s2}\, tanh(\mathbf{W}_{s1}\, \mathbf{H})), \tag{3}$$

where r is the number of global attentions we want to extract from the sentence, and $\mathbf{W}_{s1}$ and $\mathbf{W}_{s2}$ are s-by-d_h and r-by-s matrices, respectively. s is a hyperparameter governing the dimensionality, and therefore maximum rank, of the attention-producing process.

With the attention calculated, we compute an $r \times d_h$ embedding matrix, $\mathbf{M} = \mathbf{GH}$, which in essence executes r weighted sums across the T hidden states, aggregating them together into r representations. Each row of $\mathbf{G}$, denoted $\mathbf{g}^i$ ($i \in \{1 \ldots r\}$), indicates how much each hidden state contributes to the final embedded representation of $\mathbf{M}$. We can thus draw a heat map for each row of the embedding matrix M (See Figure 4 for examples). This way of visualization gives hints on what is encoded in each part of the embedding, adding an extra layer of interpretation.

To provide a final global text embedding of the sentences in the report, the AETE executes max-over-r pooling across $\mathbf{M}$, producing an embedding vector $\hat{\mathbf{X}}_{AETE}$ with size d_h.

3.3. Saliency Weighted Global Average Pooling

In addition to using attention to provide a more meaningful text embedding, our goal is also to produce improved visual embeddings for classification. For this purpose, we re-use the attention mechanism, $\mathbf{G}$, except that we perform a max-over-r operation, producing a sequence of saliency values, $g_t (t = 1, \ldots, T)$, for each word, $\mathbf{w}_t$. These saliency values are used to weight and select the spatial attention maps, $\mathbf{a}_t$, generated at each time point:

$$\mathbf{a}_{ws}(x, y) = \sum_t \mathbf{a}_t(x, y) * g_t. \tag{4}$$

This map is encoded with all spatial saliency regions guided by the text attention. We use this this map to highlight the spatial regions of $\mathbf{X}$ with more meaningful information:

$$\hat{\mathbf{X}}_{SW-GAP}(c) = \sum_{(x,y)} \mathbf{a}_{ws}(x, y) * \mathbf{X}(x, y, c), \tag{5}$$

where $x, y \in \{1...D\}$ and $\hat{\mathbf{X}}_{SW-GAP}$ is a 1-by-C vector representing the global visual information, guided by both text- and visual-based attention. The lower part of figure 2 illustrates an example of such pooling strategy.

3.4. Joint Learning

With global representations computed for both the image and report, these must be combined together to produce the final classification. To accomplish this, we concatenate the two forms of representations $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_{AETE}; \hat{\mathbf{X}}_{SW-GAP}]$ and use a final fully-connected layer to produce the output for multi-label classification. The intuition behind our model is that the connection between the CNN and RNN network will benefit the training of both because the image activations can be adjusted for the text embedding task and salient image features could be extracted by pooling based on high text saliency.

In a similar fashion as Wang *et al.* [35], we define an M-dimensional disease label vector $\mathbf{y} = [y_1, ..., y_m, ..., y_M], y_m \in \{0, 1\}$ for each case and $M = 15$ indicates the number of classes. y_m indicates the presence with respect to a pathology or 'no finding' (of listed disease categories) in the image. Here, we adopt the NLP-mined labels provided by [35] as the 'ground-truth' during the training.

The instance numbers for different disease categories are highly unbalanced, from hundreds to dozens of thousands. In addition to the positive/negative balancing introduced in [35], we add weights to instances associated with different categories,

$$\begin{aligned} L_m(f(I, \mathbf{S}), \mathbf{y}) = \beta_P \sum_{y_m=1} - \ln(f(I, \mathbf{S})) \cdot \lambda_m \\ + \beta_N \sum_{y_m=0} - \ln(1 - f(I, \mathbf{S})) \cdot \lambda_m, \end{aligned} \quad (6)$$

where $\beta_P = \frac{|N|}{|P|+|N|}$ and $\beta_N = \frac{|P|}{|P|+|N|}$. $|P|$ and $|N|$ are the total number of images with at least one disease and with no diseases, respectively. $\lambda_m = (Q - Q_m)/Q$ is a set of precomputed class-wised weights, where Q and Q_m are the total number of images and the number of images that have disease label m. λ_m will be larger if the number of instances from class m is small.

Because the TieNet can also generate text reports, we also optimize the RNN generative model loss [37], L_R. Thus the overall loss is composed of two parts, the sigmoid cross entropy loss L_C for the multi-label classification and the loss L_R from the RNN generative model [37],

$$L_{overall} = \alpha L_C + (1 - \alpha) L_R \quad (7)$$

where α is added to balance the large difference between the two loss types.

3.5. Medical Image Auto-Annotation

One straightforward application of the TieNet is the auto-annotation task to mine image classification labels. By omitting the generation of sequential words, we accumulate and back-propagate only the classification loss for better text-image embeddings in image classification. Here, we use the NLP-mined disease labels as 'ground truth' in the training. Indeed we want to learn a mapping between the input image-report pairs and the image labels. The report texts often contain more easy-to-learn features than the image side. The contribution of both sources to the final classification prediction should be balanced via either controlling the feature dimensions or drop-off partial of the 'easy-to-learn' data during training.

3.6. Automatic Classification and Reporting of Thorax Diseases

For a more difficult but real-world scenario, we transform the image-text embedding network to serve as a unified system of image classification and report generation when only the unseen image is available. During the training, both image and report are fed and two separate losses are computed as stated above, *i.e.*, the loss for image classification and the loss for sequence-to-sequence modeling. While testing, only the image is required as the input. The generated text contained the learned text embedding recorded in the LSTM units and later used in the final image classification task. The generative model we integrated into the text-image embedding network is the key to associate an image with its attention encoded text embedding.

4. Dataset

ChestX-ray14 [35] is a recently released benchmark dataset for common thorax disease classification and localization. It consists of 14 disease labels that can be observed in chest X-ray, *i.e.*, Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. The NLP-mined labels are used as 'ground truth' for model training and testing throughout the experiments. We adopt the patient-level data splits published with the data [1].

Hand-labeled: In addition to NLP-mined labels, we randomly select 900 reports from the testing set and have two radiologists to annotate the 14 categories of findings for the evaluation purpose. A trial set of 30 reports was first used to synchronize the criterion of annotation between two annotators. Then, each report was independently annotated by two annotators. In this paper, we used the inter-rater agreement (IRA) to measure the consistency between two observers. The resulting Cohens kappa is 84.3%. Afterwards, the final decision was adjudicated between two observers on the inconsistent cases.

OpenI [7] is a publicly available radiography dataset collected from multiple institutes by Indiana University. Using the OpenI API, we retrieved 3,851 unique radiology reports and 7,784 associated frontal/lateral images where each OpenI report was annotated with key concepts (MeSH words) including body parts, findings, and diagnoses. For consistency, we use the same 14 categories of findings as above in the experiments. In our experiments, only 3,643 unique front view images and corresponding reports are selected and evaluated.

5. Experiments

Report vocabulary: We use all 15,472 unique words in the training set that appear at least twice. Words that appear less frequently are replaced by a special out-of-vocabulary token, and the start and the end of the reports are marked with a special ⟨START⟩ and ⟨END⟩ token. The pre-trained

[1] https://nihcc.app.box.com/v/ChestXray-NIHCC

word embedding vectors was learned on PubMed articles using the gensim word2vec implementation with the dimensionality set to 200 [2]. The word embedding vectors will be evolved along with other LSTM parameters.

Evaluation Metrics: To compare previous state-of-the-art works, we choose different evaluation metrics for different tasks so as to maintain consistency with data as reported in the previous works.

Receiver Operating Curves (ROC) are plotted for each disease category to measure the image classification performance and afterward, Areas Under Curve (AUC) are computed, which reflect the overall performance as a summary of different operating points.

To assess the quality of generated text report, BLEU scores [28], METEOR [3] and ROUGE-L [20] are computed between the original reports and the generated ones. Those measures reflect the word overlapping statistics between two text corpora. However, we believe their capabilities are limited for showing the actual accuracy of disease words (together with their attributes) overlapping between two text corpora.

Training: The LSTM model contains a 256 dimensional cell and $s = 2000$ in $\mathbf{W}_{s1}$ and $\mathbf{W}_{s2}$ for generating the attention weights $\mathbf{G}$. During training, we use 0.5 dropout on the MLP and 0.0001 for L2 regularization. We use the Adam optimizer with a mini-batch size of 32 and a constant learning rate of 0.001.

In addition, our self-attention LSTM has a hidden layer with 350 units. We choose the matrix embedding to have 5 rows (the r), and a coefficient of 1 for the penalization term. All the models are trained until convergence is achieved and the hyper-parameters for testing is selected according to the corresponding best validation set performance.

Our text-image embedding network is implemented based on TensorFlow [1] and Tensorpack [3]. The ImageNet pre-trained model, *i.e.*, ResNet-50 [12] is obtained from the Caffe model zoo and converted into the TensorFlow compatible format. The proposed network takes the weights from the pre-trained model and fixes them during the training. Other layers in the network are trained from scratch. In a similar fashion as introduced in [35], we reduce the size of mini-batch to fit the entire model in each GPU while we accumulate the gradients for a number of iterations and also across a number of GPUs for better training performance. The DCNN models are trained using a Dev-Box Linux server with 4 Titan X GPUs.

5.1. Auto-annotation of Images

Figure 3 illustrates the ROC curves for the image classification performance with 3 different inputs evaluated on 3 different testing sets, *i.e.*, ChestX-ray14 testing set (ChestX-ray14), the hand-labeled set (Hand-labeled) and the OpenI set (OpenI). Separate curves are plotted for each disease categories and 'No finding'. Here, two different auto-annotation frameworks are trained by using different inputs, *i.e.*, taking reports only (R) and taking image-report pairs (I+R) as inputs. When only the reports are used, the framework will not have the saliency weighted global average pooling path. In such way, we can get a sense how the features from text path and image path individually contribute to the final classification prediction.

We train the proposed auto-annotation framework using the training and validation sets from the ChestX-ray14 dataset and test it on all three testing sets, *i.e.*, ChestX-ray14, hand-labeled and OpenI. Table 1 shows the AUC values for each class computed from the ROC curves shown in Figure 3. The auto-annotation framework achieves high performance on both ChestX-ray14 and Hand-labeled, *i.e.*, over 0.87 in AUC with reports alone as the input and over 0.90 in AUC with image-report pairs on sample number weighted average ($\#wAVG$). The combination of image and report demonstrates the supreme advantage in this task. In addition, the auto-annotation framework trained on ChestX-ray14 performed equivalently on OpenI. It indicates that the model trained on a large-scale image dataset could easily be generalized to the unseen data from other institutes. The model trained solely based on images could also be generalized well to the datasets from other sources. In this case, both the proposed method and the one in [35] are able to perform equally well on all three testing sets.

5.2. Classification and Reporting of Chest X-ray

When the TieNet is switched to an automatic disease classification and reporting system, it takes a single image as the input and is capable of outputting a multi-label prediction and corresponding radiological report together. The ROC curves on the right in Figure 3 and Table 1 show the image classification performance produced by the multi-purpose reporting system. The AUCs from our TieNet (I+GR) demonstrate the consistent improvement in AUCs ($2.3\% - 5.7\%$ on $\#wAVG$ for all the disease categories) across all three datasets. The multilabel classification framework [35] serves as a baseline model that also takes solely the images. Furthermore, the performance improvement achieved on the Hand-labeled and OpenI datasets (with ground truth image labels) is even larger than the performance gain on ChestX-ray14 (with NLP-mined labels). It indicates that the TieNet is able to learn more meaningful and richer text embeddings directly from the raw reports and correct the inconsistency between embedded features and erroneous mined labels.

Table 2 shows that the generated reports from our proposed system obtain higher scores in all evaluation metrics

[2] https://radimrehurek.com/gensim/models/word2vec.html

[3] https://github.com/ppwwyyxx/tensorpack/

Disease	ChestX-ray14					Hand-labeled					OpenI				
	R /	I+R /	I [35] /	I+GR /	#	R /	I+R /	I [35] /	I+GR /	#	R /	I+R /	I [35] /	I+GR /	#
Atelectasis	.983 /	.993 /	.700 /	.732 /	3255	.886 /	.919 /	.680 /	.715 /	261	.981 /	.976 /	.702 /	.774 /	293
Cardiomegaly	.978 /	.994 /	.810 /	.844 /	1065	.964 /	.989 /	.820 /	.872 /	185	.944 /	.962 /	.803 /	.847 /	315
Effusion	.984 /	.995 /	.759 /	.793 /	4648	.938 /	.967 /	.780 /	.823 /	257	.968 /	.977 /	.890 /	.899 /	140
Infiltration	.960 /	.986 /	.661 /	.666 /	6088	.849 /	.879 /	.648 /	.664 /	271	.981 /	.984 /	.585 /	.718 /	57
Mass	.984 /	.994 /	.693 /	.725 /	1712	.935 /	.943 /	.696 /	.710 /	93	.959 /	.903 /	.756 /	.723 /	14
Nodule	.981 /	.994 /	.668 /	.685 /	1615	.974 /	.974 /	.662 /	.684 /	130	.967 /	.960 /	.647 /	.658 /	102
Pneumonia	.947 /	.969 /	.658 /	.720 /	477	.917 /	.946 /	.724 /	.681 /	55	.983 /	.994 /	.642 /	.731 /	36
Pneumothorax	.983 /	.995 /	.799 /	.847 /	2661	.983 /	.996 /	.784 /	.855 /	166	.960 /	.960 /	.631 /	.709 /	22
Consolidation	.989 /	.997 /	.703 /	.701 /	1815	.923 /	.910 /	.609 /	.631 /	60	.969 /	.989 /	.790 /	.855 /	28
Edema	.976 /	.989 /	.805 /	.829 /	925	.970 /	.987 /	.815 /	.834 /	33	.984 /	.995 /	.799 /	.879 /	40
Emphysema	.996 /	.997 /	.833 /	.865 /	1093	.980 /	.981 /	.835 /	.863 /	44	.849 /	.868 /	.675 /	.792 /	94
Fibrosis	.986 /	.986 /	.786 /	.796 /	435	.930 /	.989 /	.688 /	.714 /	11	.985 /	.960 /	.744 /	.791 /	18
PT	.988 /	.997 /	.684 /	.735 /	1143	.904 /	.923 /	.679 /	.776 /	41	.948 /	.953 /	.691 /	.749 /	52
Hernia	.929 /	.958 /	.871 /	.876 /	86	.757 /	.545 /	.864 /	.647 /	2	– /	– /	– /	– /	0
NoFinding	.920 /	.985 /	– /	.701 /	9912	.889 /	.908 /	– /	.666 /	85	.933 /	.936 /	– /	.747 /	2789
AVG	.976 /	.989 /	.745 /	.772 /	–	.922 /	.925 /	.735 /	.748 /	–	.960 /	.965 /	.719 /	.779 /	–
#wAVG	.978 /	.992 /	**.722** /	**.748** /	–	.878 /	.900 /	**.687** /	**.719** /	–	.957 /	.966 /	**.741** /	**.798** /	–

Table 1. Evaluation of image classification results (AUCs) on ChestX-ray14, hand-labeled and OpenI dataset. Performances are reported on four methods, *i.e.*, multilabel classification based on Report (R), Image + Report (I+R), Image [35], and Image + Generative Report(I+GR).

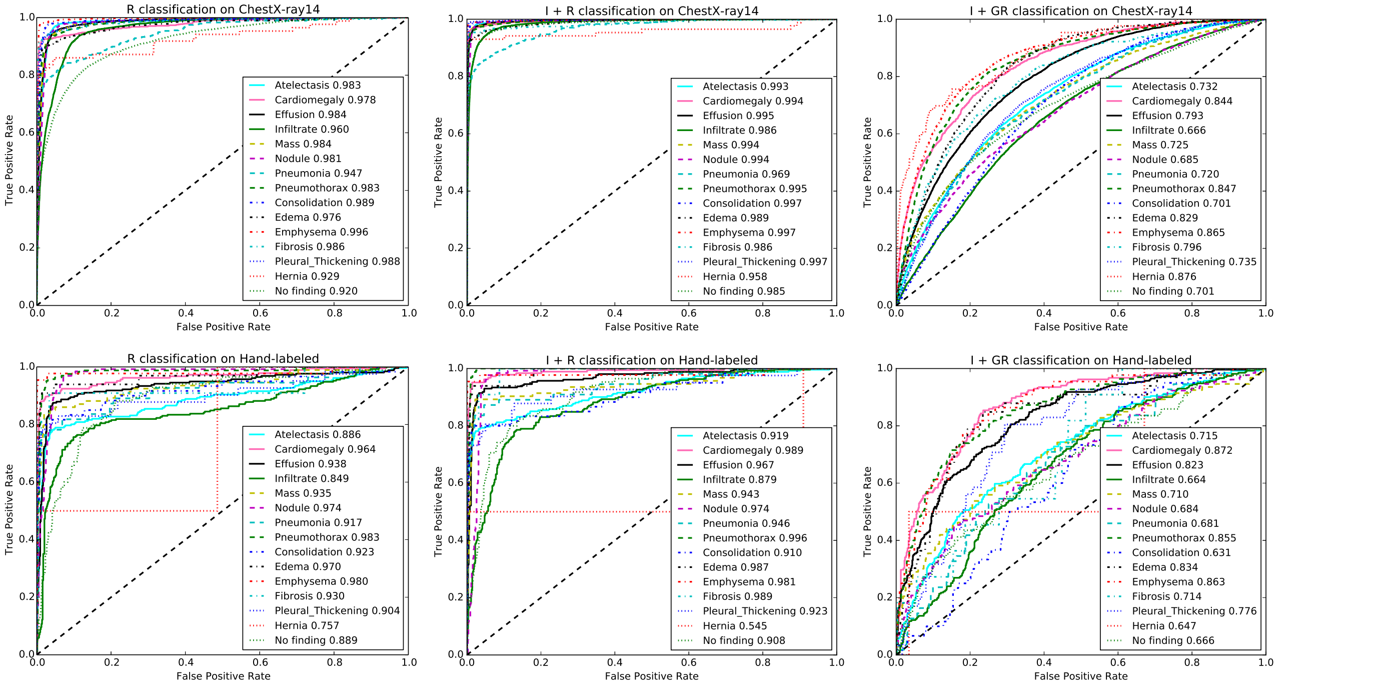
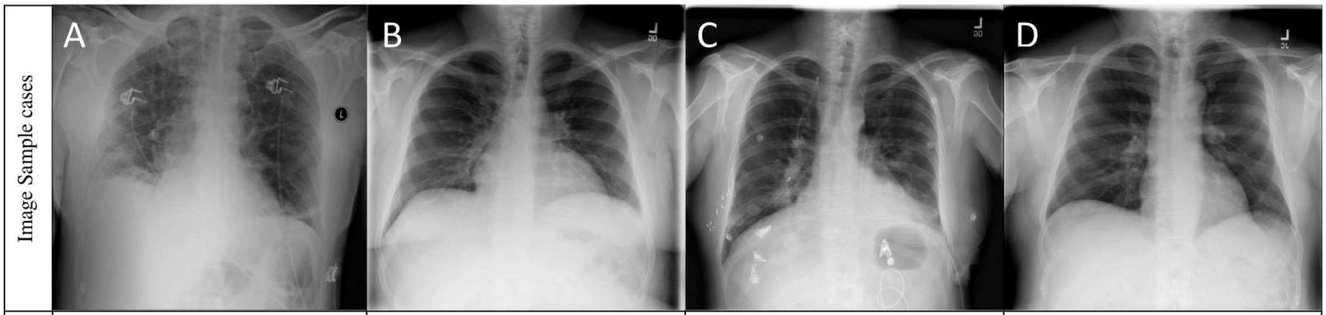


Figure 3. A comparison of classification performance using ROC curves with different testing inputs, *i.e*. Report (R), Image+Report (I+R), and Image+Generative Report(I+GR).

in comparison to the baseline image captioning model [37]. It may be because the gradients from RNN are backpropagated to the CNN part and the adjustment of image features from Transition layer will benefit the report generation task.

Figure 4 illustrates 4 sample results from the proposed automatic classification and reporting system. Original images are shown along with the classification predications, original reports and generated reports. Text-attended words



	A	B	C	D
P	Atelectasis Effusion	No finding	Nodule Pneumothorax Mass Consolidation	Mass
Original report	findings : a single ap view of the chest demonstrates increasing bibasilar interstitial opacities with decreased overall aeration . increasing blunting of right costophrenic angle. … impression : increasing bibasilar atelectasis with possible development of right pleural effusion .	Normal no evidence of lung infiltrate .	findings : heart and mediastinum unchanged . multiple lung nodules . evidence of recent left chest surgery with left chest tube in place . very small left apical pneumothorax . lungs unchanged , no evidence of acute infiltrates . impression : stable chest .	findings : large left suprahilar and infrahilar masses as well as the well circumscribed nodule the level of the aortic knob . the right infrahilar mass as well . no effusion . impression : metastatic lung disease .
Generated Report	findings : a single ap view of the chest demonstrates unchanged bilateral reticular opacities , consider atelectasis . continued left basilar atelectasis . no evidence of developing infiltrate . the cardiac and mediastinal contours are stable . impression : no evidence of developing infiltrate .	findings : pa and lateral views of the chest demonstrate lungs that are clear without focal mass , infiltrate or effusion . cardiomediastinal silhouette is normal size and contour . pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology	findings : pa and lateral views of the chest demonstrate unchanged bilateral chest tubes . again pulmonary nodules are seen on the right and cardiac silhouette unchanged . the cardiac and mediastinal contours are stable . impression : 1. bilateral masses and left lower lung field consolidation . 2.new bilateral lung masses .	comparison is to previous upright study of no significant interval change is seen in the appearance of the chest . the mediastinal soft tissue and pulmonary vascularity are stable . there are blastic bone lesions in the chest . bones , soft tissues are normal . the lung fields are clear . there are calcified lymph nodes in the left lower lung . impression : . sclerotic lesions in the left humeral , consistent with metastasis.

Figure 4. 4 sample image Classification Predictions (P) along with original and generated reports. Text attentions are highlighted over the generated text. Correct predication is marked in green, false prediction in red and missing prediction in blue.

Table 2. Evaluation of generated reports in ChestX-ray14 testing set using BLEU, METEOR and ROUGE-L.

	Captioning [37]	TieNet I+GR
BLEU-1	0.2391	0.2860
BLEU-2	0.1248	0.1597
BLEU-3	0.0861	0.1038
BLEU-4	0.0658	0.0736
METEOR	0.1024	0.1076
ROUGE-L	0.1988	0.2263

are also highlighted over the generated reports. If looking at generated reports alone, we find that they all read well. However, the described diseases may not truly appear in the images. For example, 'Atelectasis' is correctly recognized in sample A but 'Effusion' is missed. 'Effusion' (not too far from the negation word 'without') is erroneously highlighted in sample B but the system is still able to correctly classify the image as 'No finding'. In sample D, the generated report misses 'Mass' while it states right about the metastasis in the lung. One promising finding is that the false predictions ('Mass' and 'Consolidation') in sample C can actually be observed in the image (verified by a radiologist) but somehow were not noted in the original report, which indicates our proposed netowrk can in some extent associate the image appearance with the text description.

6. Conclusion

Automatically extracting the machine-learnable annotation from the retrospective data remains a challenging task, among which images and reports are two main useful sources. Here, we proposed a novel text-image embedding network integrated with multi-level attention models. TieNet is implemented in an end-to-end CNN-RNN architecture for learning a blend of distinctive image and text representations. Then, we demonstrate and discuss the pros and cons of including radiological reports in both auto-annotation and reporting tasks. While significant improvements have been achieved in multi-label disease classification, there is still much space to improve the quality of generated reports. For future work, we will extend TieNet to include multiple RNNs for learning not only disease words but also their attributes and further correlate them and image findings with the description in the generated reports.

Acknowledgements This work was supported by the Intramural Research Programs of the NIH Clinical Center and National Library of Medicine. Thanks to Adam Harrison and Shazia Dharssi for proofreading the manuscript. We are also grateful to NVIDIA Corporation for the GPU donation.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 2016. 6

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, pages 1–15, 2015. 3

[3] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 6

[4] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai, and H. Greenspan. Fully convolutional network for liver segmentation and lesions detection. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 77–85, 2016. 2

[5] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang. Deep learning: a primer for radiologists. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 37(7):2113–2131, 2017. 2

[6] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3086, 2017. 1

[7] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015. 2, 5

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 2

[10] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 1

[11] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–13, 2017. 1

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 4

[14] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1115–1124, 2017. 1

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014. 1

[16] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016. 1

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):664–676, 2017. 1

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1

[19] 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

[20] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, pages 1–8. Barcelona, Spain, 2004. 6

[21] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations (ICLR)*, pages 1–10, 2014. 1

[22] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 1

[23] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations (ICLR)*, pages 1–15, 2017. 3, 4

[24] J. Liu, D. Wang, L. Lu, Z. Wei, L. Kim, E. B. Turkbey, B. Sahiner, N. Petrick, and R. M. Summers. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Medical Physics*, 44(9):4630–4642, 2017. 2

[25] Y. Liu, C. Sun, L. Lin, and X. Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. 2016. 3

[26] F. Meng, Z. Lu, M. Wang, H. Li, W. Jiang, and Q. Liu. Encoding source language with convolutional neural network

for machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-CoNLL)*, pages 20–30, 2015. 3

[27] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 299–307, 2017. 1, 2

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics (ACL)*, pages 311–318, 2002. 6

[29] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *International Conference on Computer Vision (ICCV)*, pages 1–22, 2017. 2

[30] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[31] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389, 2015. 3

[32] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to read chest X-rays: recurrent neural cascade model for automated image annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506, 2016. 2

[33] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015. 2

[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 1

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017. 2, 3, 5, 6, 7

[36] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: free-form visual question answering based on knowledge from external sources. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–5, 2016. 1

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. 2, 4, 5, 7, 8

[38] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level attention networks for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2017. 1, 2

[39] W. L. L. C.-C. Yulia, T. S. Amir, R. F. A. C. D. Alan, and W. B. I. Trancoso. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1367–1372, 2015. 3

[40] Z. Zhang, P. Chen, M. Sapkota, and L. Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–328. Springer, 2017. 2

[41] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. MDNet: a semantically and visually interpretable medical image diagnosis network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6428–6436, 2017. 2

[42] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded question answering in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1