

On the Automatic Generation of Medical Imaging Reports

Baoyu Jing Pengtao Xie Eric Xing
Petuum Inc.

{baoyu.jin, pengtao.xie, eric.xing}@petuum.com

Abstract

Medical imaging is widely used in clinical practice for diagnosis and treatment. Specialized physicians read medical images and write textual reports to narrate the findings. Report-writing can be error-prone for unexperienced physicians, and time-consuming and tedious for physicians in highly populated nations. To address these issues, we study the automatic generation of medical imaging reports, as an assistance for human physicians in producing reports more accurately and efficiently. This task presents several challenges. First, a complete report contains multiple heterogeneous forms of information, including findings which are paragraphs and tags which are a list of key words. Second, abnormal regions in medical images are difficult to identify. Generating textual narrations for them is even harder. Third, the reports are typically long, containing multiple paragraphs. To cope with these challenges, we (1) build a multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs, (2) propose a co-attention mechanism to localize regions containing abnormalities and generate narrations for them, (3) develop a hierarchical LSTM model to generate long paragraphs. We demonstrate the effectiveness of the proposed methods on a chest x-ray dataset and a pathology dataset.

1. Introduction

Medical images, such as radiology and pathology images, are widely used in hospitals and clinics for the diagnosis and treatment of many diseases, such as pneumonia, pneumothorax, interstitial lung disease, heart failure, bone fracture, hiatal hernia, to name a few. The reading and interpretation of medical images are usually conducted by specialized medical professionals. For example, radiology and pathology images are read by radiologists and pathologists. They write textual reports (Figure 1) to narrate the findings regarding each area of the body examined in the imaging study, specifically whether each area was found to be normal, abnormal or potentially abnormal.

For less-experienced radiologists and pathologists, es-



Impression: No acute cardiopulmonary abnormality.

Findings: There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pneumothorax. Degenerative changes of the thoracic spine.

MTI Tags: degenerative change

Figure 1. An exemplar chest x-ray report, which consists of multiple sections of information. In the *impression* section, the radiologist combines the findings, patient clinical history and indication for the imaging study and provides a diagnosis. The *findings* section lists the radiology observations and findings regarding each area of the body examined in the imaging study. The *tags* section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer (MTI).

pecially those working in the rural area where the quality of healthcare is relatively low, writing medical-imaging reports is demanding. For instance, to correctly read a chest x-ray image, the following skills are needed [2]: (1) thorough knowledge of the normal **anatomy** of the **thorax**, and the basic physiology of chest diseases; (2) skills of analyzing the radiograph through a fixed pattern; (3) ability of evaluating the evolution over time; (4) knowledge of clinical presentation and history; (5) knowledge of the correlation with other diagnostic results (laboratory results, electrocardiogram, respiratory function tests).

For experienced radiologists and pathologists, writing imaging reports is tedious and time-consuming. In nations with large population such as China, a radiologist may need to read hundreds of radiology images per day. Typing the findings of each image into computer takes about 5-10 minutes, which occupies most of their working time. In sum, for both unexperienced and experienced medical professionals, writing imaging reports is unpleasant.

This motivates us to investigate whether it is possible to automatically generate medical image reports. Several challenges need to be addressed. First, a complete diagnostic report is comprised of multiple heterogeneous forms of in-

formation. As shown in Figure 1, the report for a chest x-ray contains *impression* which is a sentence, *findings* which are a paragraph and *tags* which are a list of keywords. Generating this heterogeneous information in a unified framework is technically demanding. We address this problem by building a multi-task framework, which treats the prediction of tags as a multi-label classification task, and treats the generation of long descriptions (e.g. impressions, findings) as a text generation task. In the framework, the two tasks share the same CNN used for learning visual features and are performed jointly. Secondly, an imaging report usually focuses more on narrating the abnormal findings since they directly indicate diseases and guide treatment. How to localize image-regions that contain abnormalities and attach the right description to them are challenging. We solve this problem by introducing a co-attention mechanism, which simultaneously attends to images and predicted tags and explores the synergistic effects of visual and semantic information. Third, the descriptions in imaging reports are usually long, containing multiple sentences or even multiple paragraphs. Generating such long text is highly non-trivial. Rather than adopting a single-layer LSTM, which is less capable of modeling long word sequences, we leverage the compositional nature of the report and adopt a hierarchical LSTM to produce long texts. Combined with the co-attention mechanism, the hierarchical LSTM first generates high-level topics, and then produces fine-grained descriptions according to the topics.

Overall, the main contributions of our work are:

- We propose a multi-task learning framework which can simultaneously predict the tags and generate the text descriptions.
- We introduce a co-attention mechanism for localizing abnormal regions and generating the corresponding descriptions.
- We build a hierarchical LSTM to generate long sentences and paragraphs.
- We perform extensive quantitative and qualitative experiments to show the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 introduces the method. Section 4 present the experimental results and Section 5 concludes the paper.

2. Related works

Textual labeling of medical images There have been several works aiming at attaching “texts” to medical images. In their settings, the target “texts” are either fully-structured or semi-structured (e.g. tags, attributes, templates), rather

than natural texts. Kisilev et al. [10] build a pipeline to predict the attributes of medical images. Some of these attributes are textual tags. Given a medical image, they first perform image segmentation, then extract visual features of the segments and finally build a classifier to categorize the segments into predefined (textual) categories. Shin et al. [16] adopt a CNN-RNN based framework to predict tags of chest x-ray images. They use CNN to detect a disease from an image and use RNN to describe the contexts (e.g., location, severity and the affected organs) of the detected disease. The work closest to *medical report generation* is recently contributed by Zhang et al. [22]. They aim at generating pathology reports containing 30-59 words. However, the reports studied in their work are semi-structured and are not fully natural. Most reports are generated by rephrasing a small number of “standard” reports, whose contents are restricted to 5 predefined topics.

To our best knowledge, our work represents the first one that generates truly natural reports written down by physicians. These reports are long and cover diverse topics, which are much more challenging to generate than tags and semi-structured paragraphs studied in previous works.

Image captioning with deep learning Image captioning aims at automatically generating text descriptions for given images. Most of recent image captioning models are based on a CNN-RNN framework [19, 6, 9, 20, 21, 8, 11]. Vinyals et al. [19] feed the image features extracted from the last hidden layer of CNN into an LSTM to generate captions. Fang et al. [6] first use CNN to detect the concepts in the image, and then use these detected concepts to guide the language model to generate a sentence. Karpathy et al. [9] propose to use multi-modal recurrent neural network to learn the alignment between visual and semantic features, and then produce the descriptions for given images.

Recently, attention mechanisms have shown to be useful for image captioning [20, 21]. Xu et al. [20] introduce a spatial-visual attention mechanism over image features extracted from intermediate layers of CNN. You et al. [21] propose a semantic attention mechanism over tags of given images. To better leverage both of the visual features and semantic tags, we propose a co-attention mechanism for report generation.

Instead of only generating one sentence caption for an image, Johnson et al. [8] introduce the dense captioning task which requires the model to generate a description for each of the detected image regions. Krause et al. [11] and Liang et al. [12] generate paragraph captions for images through hierarchical LSTM. Our method also adopts a hierarchical LSTM for paragraph generation, whereas different from [11], we use a co-attention network to generate topics.

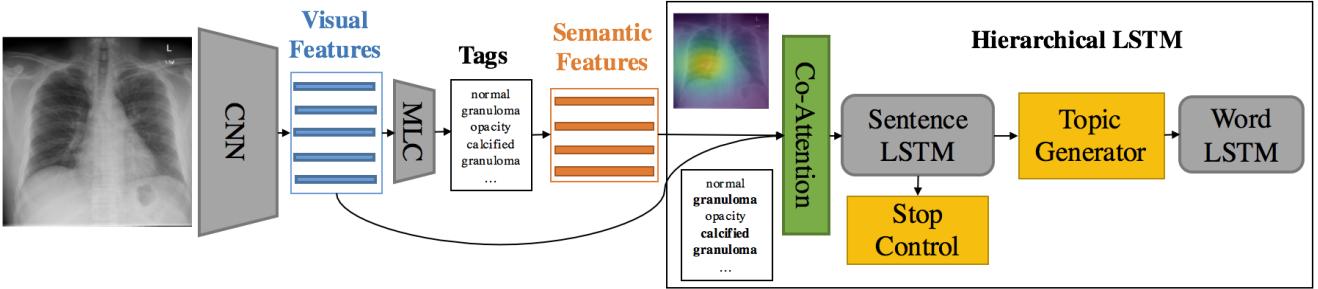


Figure 2. Illustration of the proposed model. MLC denotes an *multi-label classification* network. Semantic features are the word embeddings of the predicted tags. The boldfaced tags “calcified granuloma” and “granuloma” are attended by the co-attention network.

3. Methods

In this section, we introduce the methods, beginning with an overview followed by the detailed description of each component.

3.1. Overview

A complete diagnostic report for medical image is comprised of both unstructured descriptions (in the form of sentences and paragraphs) and semi-structured tags (in the form of keyword lists), as shown in Figure 1. We propose a *multi-task hierarchical model with co-attention* for automatically predicting keywords and generating long paragraphs. Given an image which is divided into regions, we use a CNN to learn visual features for these patches. Then these visual features are fed into a *multi-label classification* (MLC) network to predict the relevant tags. In the tag vocabulary, each tag is represented by a word-embedding vector. Given the predicted tags for a specific image, their word-embedding vectors are retrieved to serve as the semantic features of this image. Then the visual features and semantic features are fed into a *co-attention model* to generate a context vector that simultaneously captures the visual and semantic information of this image. As of now, the encoding process is completed. Next, starting from the context vector, the decoding process generates the text descriptions. The description of an medical image usually contains multiple sentences, and each sentence focuses on one specific topic. Our model leverages this compositional structure to generate reports in a hierarchical way: it first generates a sequence of high-level topic vectors representing sentences, then generates a sentence (a sequence of words) from each topic vector. Specifically, the context vector is inputted into a *sentence LSTM*, which unrolls for a few steps, each producing a topic vector. A topic vector represents the semantics of a sentence to be generated. Given a topic vector, the *word LSTM* takes it as input and generates a sequence of words to form a sentence. The termination of the unrolling process is controlled by the sentence LSTM.

3.2. Tag prediction

The first task of our model is predicting the tags of the given image. We treat the tag prediction task as a multi-label classification task. Specifically, given an image I , we first extract its features $\{\mathbf{v}_n\}_{n=1}^N \in \mathbb{R}^D$ from an intermediate layer of a CNN, and then feed $\{\mathbf{v}_n\}_{n=1}^N$ into a *multi-Label classification* (MLC) network to generate a distribution over all of the L tags:

$$\mathbf{p}_{i,pred}(\mathbf{l}_i = 1 | \{\mathbf{v}_n\}_{n=1}^N) \propto \exp(\text{MLC}_i(\{\mathbf{v}_n\}_{n=1}^N)) \quad (1)$$

where $\mathbf{l} \in \mathbb{R}^L$ is tag vector, $\mathbf{l}_i = 1$ and $\mathbf{l}_i = 0$ denote the presence and absence of the i -th tag respectively, and MLC_i means the i -th output of MLC network.

For simplicity, we extract visual features from the last convolutional layer of the VGG-19 [17] and adopt the last two fully connected layers of VGG-19 as MLC.

Finally, the embedding of the top M tags $\{\mathbf{a}_m\}_{m=1}^M \in \mathbb{R}^E$ are used as semantic features for topic generation.

3.3. Co-Attention

Previous works have shown that visual attention alone can perform fairly well for localizing the objects [1] and aiding caption generation [20]. However, visual attention doesn't provide sufficient high level semantic information. For example, only looking at the right lower region of the chest x-ray image (Figure 1) without concerning about other areas, we might not be able to recognize what we are looking at, not to even mention detecting the abnormalities. This will be problematic for topic generation. On the contrary, the tags can always provide the needed high level information about overall or part of images. To this end, we propose a *co-attention mechanism* which can simultaneously attend to visual and semantic modalities.

At time step s of sentence LSTM, the joint context vector $\mathbf{ctx}^{(s)} \in \mathbb{R}^C$ is generated by co-attention network $f_{co_{att}}(\{\mathbf{v}_n\}_{n=1}^N, \{\mathbf{a}_m\}_{m=1}^M, \mathbf{h}_{sent}^{(s-1)})$, where $\mathbf{h}_{sent}^{(s-1)} \in \mathbb{R}^H$ is the hidden state of sentence LSTM at time step $s - 1$. The co-attention network $f_{co_{att}}$ first adopts a single layer feed-forward network to compute the soft visual and semantic

attentions over input image features and tags:

$$\alpha_{v,n} \propto \exp(\mathbf{W}_{v,att} \tanh(\mathbf{W}_v \mathbf{v}_n + \mathbf{W}_{v,h} \mathbf{h}_{sent}^{(s-1)})) \quad (2)$$

$$\alpha_{a,m} \propto \exp(\mathbf{W}_{a,att} \tanh(\mathbf{W}_a \mathbf{a}_m + \mathbf{W}_{a,h} \mathbf{h}_{sent}^{(s-1)})) \quad (3)$$

where \mathbf{W}_v , $\mathbf{W}_{v,h}$ and $\mathbf{W}_{v,att}$ are parameter matrices of the visual attention network. \mathbf{W}_a , $\mathbf{W}_{a,h}$ and $\mathbf{W}_{a,att}$ are parameter matrices of the semantic attention network.

The soft visual and semantic attentions are thus:

$$\mathbf{v}_{att}^{(s)} = \sum_{n=1}^N \alpha_{v,n} \mathbf{v}_n, \quad \mathbf{a}_{att}^{(s)} = \sum_{m=1}^M \alpha_{a,m} \mathbf{a}_m$$

There are many ways to combine the visual and semantic context vectors such as concatenation and element-wise operations. In this paper, we first concatenate visual $\mathbf{v}_{att}^{(s)}$ and semantic $\mathbf{a}_{att}^{(s)}$ context vectors as $[\mathbf{v}_{att}^{(s)}; \mathbf{a}_{att}^{(s)}]$, and then use a fully connected layer \mathbf{W}_{fc} to obtain joint context vector:

$$\mathbf{ctx}^{(s)} = \mathbf{W}_{fc} [\mathbf{v}_{att}^{(s)}; \mathbf{a}_{att}^{(s)}] \quad (4)$$

3.4. Sentence LSTM

The sentence LSTM is a single-layer LSTM that takes the joint context vector $\mathbf{ctx}^{(s)}$ as its input, and it generates topic vector $\mathbf{t} \in \mathbb{R}^K$ for word LSTM through topic generator and determines whether to continue or stop generating captions by stop control component.

Topic Generator We use a deep output layer [15] to strengthen the context information in topic vector $\mathbf{t}^{(s)}$. The deep output layer takes as input the hidden state $\mathbf{h}_{sent}^{(s)}$ and the joint context vector $\mathbf{ctx}^{(s)}$ of current step:

$$\mathbf{t}^{(s)} = \tanh(\mathbf{W}_{t,h_{sent}} \mathbf{h}_{sent}^{(s)} + \mathbf{W}_{t,ctx} \mathbf{ctx}^{(s)}) \quad (5)$$

where $\mathbf{W}_{t,h_{sent}}$ and $\mathbf{W}_{t,ctx}$ are the parameter matrices.

Stop Control We also apply a deep output layer to control the continuation of the sentence LSTM. The layer takes as input the previous and current hidden state $\mathbf{h}_{sent}^{(s-1)}$, $\mathbf{h}_{sent}^{(s)}$ and produces a distribution over {STOP=1, CONTINUE=0}:

$$p(STOP | \mathbf{h}_{sent}^{(s-1)}, \mathbf{h}_{sent}^{(s)}) \\ \propto \exp(W_{Stop} \tanh(\mathbf{W}_{Stop,s-1} \mathbf{h}_{sent}^{(s-1)} + \mathbf{W}_{Stop,s} \mathbf{h}_{sent}^{(s)})) \quad (6)$$

where \mathbf{W}_{Stop} , $\mathbf{W}_{Stop,s-1}$ and $\mathbf{W}_{Stop,s}$ are parameter matrices. If p is greater than predefined threshold (e.g. 0.5), then sentence LSTM will stop producing new topic vectors and word LSTM will also stop producing words.

3.5. Word LSTM

The words of each sentence are generated by a word LSTM, which is a single layer LSTM initialized by topic vector \mathbf{t} produced by sentence LSTM. Following [11], when generating words, topic vector \mathbf{t} and the special *START* token are used as the first and the second input of the word LSTM, and the subsequent inputs are the word sequence.

The hidden state $\mathbf{h}_{word} \in \mathbb{R}^H$ of word LSTM is directly used to predict the distribution over words by:

$$p(word | \mathbf{h}_{word}) \propto \exp(\mathbf{W}_{out} \mathbf{h}_{word}) \quad (7)$$

where \mathbf{W}_{out} is the parameter matrix. After each word-LSTM has generated its word sequences, the final report is the simple concatenation of all the generated sequences.

3.6. Parameter learning

Each training example is a tuple $(I, \mathbf{l}, \mathbf{w})$ where I is an image, \mathbf{l} denotes the ground-truth tag vector and \mathbf{w} is the diagnostic paragraph, which is comprised of S sentences, and each of sentence consists of T_s words.

Given a training example $(I, \mathbf{l}, \mathbf{w})$, our model first performs multi-label classification for I , and produces a distribution $\mathbf{p}_{l,pred}$ over all tags. Note that \mathbf{l} is a binary vector which encodes the presence and absence of tags. We can obtain ground-truth tag distribution by normalizing \mathbf{l} : $\mathbf{p}_l = \mathbf{l} / \|\mathbf{l}\|_1$. Therefore, the training loss of this step is cross-entropy loss ℓ_{tag} between \mathbf{p}_l and $\mathbf{p}_{l,pred}$.

Next, the sentence LSTM is unrolled for S steps to produce topic vectors and also distributions over {STOP, CONTINUE}: p_{stop}^s . Finally, the produced S topic vectors are fed into the word LSTM to generate words $\mathbf{w}_{s,t}$. The training loss of caption generation is the combination of two cross-entropy losses: ℓ_{sent} over stop distributions p_{stop}^s and ℓ_{word} over word distributions $p_{s,t}$.

Thus, the training loss is computed through:

$$\ell(I, \mathbf{l}, \mathbf{w}) = \lambda_{tag} \ell_{tag} + \lambda_{sent} \sum_{s=1}^S \ell_{sent}(p_{stop}^s, I\{s=S\}) \\ + \lambda_{word} \sum_{s=1}^S \sum_{t=1}^{T_s} \ell_{word}(p_{s,t}, w_{s,t}) \quad (8)$$

In addition to the above training loss, there is also a regularization loss for visual and semantic attentions. Following [20], suppose $\alpha \in \mathbb{R}^{N \times S}$ and $\beta \in \mathbb{R}^{M \times S}$ are matrices of visual and semantic attentions respectively, then the regularization loss over α and β is:

$$\ell_{re} = \lambda_{re} \left[\sum_n^N \left(1 - \sum_s^S \alpha_{n,s} \right)^2 + \sum_m^M \left(1 - \sum_s^S \beta_{m,s} \right)^2 \right] \quad (9)$$

This loss function encourages the model to pay equal attention over different image regions and different tags.

4. Experiments

4.1. Datasets

We use two publicly available medical image datasets to evaluate our proposed model.

IU X-Ray The Indiana University Chest X-Ray Collection (IU X-Ray) [3] is a set of chest x-ray images paired with their corresponding diagnostic reports. The dataset contains 7,470 pairs of images and reports. Each report consists of the following sections: *impression*, *findings*, *tags*¹, *comparison* and *indication*. In this paper, we treat the contents in *impression* and *findings* as the target captions to be generated and the MTI generated tags in *tags* as the tags of the reports (Figure 1 provides an example).

We preprocess the data through converting all tokens to lowercases, removing all of non-alpha tokens, which resulting in 572 unique tags, and 1915 unique words remaining in the entire dataset. On average, each image is associated with 2.2 tags, 5.7 sentences and each sentence contains 6.5 words. Besides, we find that top 1,000 words covers 99.0% word occurrences in the dataset, therefore we only use top 1,000 words to build dictionary. Finally, we randomly select 500 images for validation and 500 images for test.

PEIR Gross The Pathology Education Informational Resource (PEIR) digital library² is a public medical image library for medical education. We collect the images with their descriptions in Gross sub-collection of PEIR Pathology collection. The PEIR Gross dataset contains 7,442 image caption pairs from 21 different sub-categories. Different from IU X-Ray dataset, each caption in PEIR Gross dataset is a one sentence description of its corresponding image. We use this dataset to evaluate our model’s ability to generate single sentence report.

For PEIR Gross dataset, we apply the same preprocessing as IU X-Ray which yields 4,452 unique words for the dataset and 12.0 words for each image. Besides, similar to [21], we treat words with top 1000 occurrences as tags.

4.2. Implementation details

We set the dimensions of all hidden states and embeddings as 512. For words and tags, we use different embedding matrices since an tag might contain multiple words. We use full VGG-19 model [19] for tag prediction. We use the embedding of top 10 tags with highest probabilities to be the semantic feature vectors $\{\mathbf{a}_m\}_{m=1}^{M=10}$. We extract the visual features from the last convolutional layer of the same VGG-19 network, which yields a $14 \times 14 \times 512$ feature map.

¹There are two types of tags: manually generated and Medical Text Indexer (MTI) generated.

²<http://peir.path.uab.edu/library>

4.3. Baselines

We first compare our method with several state-of-the-art image captioning models: CNN-RNN [19], LRCN [5], Soft ATT [20], ATT-RK [21]. We re-implemented all of these models and adopt VGG-19 [17] as CNN encoder. However, these models are built for single sentence captions. To better show the effectiveness of hierarchical LSTM for paragraph generation, we also implement a hierarchical model without any attention: Ours-no-Attention. The input of Ours-no-Attention is the overall image feature of VGG-19 network, which has a dimension of 4096. Ours-no-Attention can be viewed as CNN-RNN [19] equipped with hierarchical LSTM decoder. Besides, to show the effectiveness of the proposed co-attention mechanism, we also implement two ablated versions of our model: Ours-Semantic-only and Ours-Visual-only, which takes only the semantic attention or visual attention context vector to produce topic vectors.

4.4. Quantitative results

We report the paragraph generation (upper part of Table 1) and one sentence generation (lower part of Table 1) results using the following captioning evaluation tools: BLEU [14], METEOR [4], ROUGE [13] and CIDEr [18].

For paragraph generation, as shown in the upper part of Table 1, it is clear that models with single LSTM decoder perform much worse than hierarchical LSTM decoder. Note that the only difference between Ours-No-Attention and CNN-RNN [19] in Table 1 is that Ours-No-Attention adopts a hierarchical LSTM decoder while CNN-RNN [19] only adopts a single-layer LSTM. The comparison between these two models directly demonstrates the effectiveness of hierarchical LSTM. This result is not surprising since it is well-known that single-layer LSTM cannot effectively model long sequences. Additionally, employing semantic attention alone (Ours-Semantic-Only) or visual attention alone (Ours-Visual-Only) to generate topic vectors does not seem to help caption generation a lot. The potential reason might be that visual attention can only capture the visual information of sub-regions of the image while it is unable to correctly describe it. While semantic attention only knows the potential abnormalities but it cannot confirm its findings by looking at images. Finally, our full model (Ours-CoAttention) achieves the best results on all of the evaluation metrics, which demonstrates the effectiveness of proposed co-attention mechanism.

For the results of single sentence generation (shown in the lower part of Table 1), the ablated versions of our model (Ours-Semantic-Only and Ours-Visual-Only) achieve competitive scores compared with the state-of-the-arts models. Our full model (Ours-CoAttention) outperforms all of the baseline models, which indicates the effectiveness of the proposed co-attention mechanism.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDER
IU X-Ray	CNN-RNN [19]	0.316	0.211	0.140	0.095	0.159	0.267	0.111
	LRCN[5]	0.369	0.229	0.149	0.099	0.155	0.278	0.190
	Soft ATT[20]	0.399	0.251	0.168	0.118	0.167	0.323	0.302
	ATT-RK[21]	0.369	0.226	0.151	0.108	0.171	0.323	0.155
	Ours-no-Attention	0.505	0.383	0.290	0.224	0.200	0.420	0.259
	Ours-Semantic-only	0.504	0.371	0.291	0.230	0.207	0.418	0.286
	Ours-Visual-only	0.507	0.373	0.297	0.238	0.211	0.426	0.300
	Ours-CoAttention	0.517	0.386	0.306	0.247	0.217	0.447	0.327
PEIR Gross	CNN-RNN[19]	0.247	0.178	0.134	0.092	0.129	0.247	0.205
	LRCN[5]	0.261	0.184	0.136	0.088	0.135	0.254	0.203
	Soft ATT[20]	0.283	0.212	0.163	0.113	0.147	0.271	0.276
	ATT-RK[21]	0.274	0.201	0.154	0.104	0.141	0.264	0.279
	Ours-Semantic-only	0.263	0.191	0.145	0.098	0.138	0.261	0.274
	Ours-Visual-only	0.284	0.209	0.156	0.105	0.149	0.274	0.28
	Ours-CoAttention	0.300	0.218	0.165	0.113	0.149	0.279	0.329

Table 1. Main results for paragraph generation on IU X-Ray dataset (upper part), and single sentence generation on PEIR Gross dataset (lower part). BLUE-n denotes the BLEU score uses up to n-grams.

4.5. Qualitative Results

4.5.1 Paragraph Generation

An illustration of paragraph generation by three models: Ours-CoAttention, Ours-No-Attention and Soft Attention models is shown in Figure 3. Note that the underlined sentences are the descriptions of abnormalities. First property we can observe is that the generated paragraphs are comprised of more sentences than the ground truth. Secondly, most of the generated sentences and ground truth sentences are descriptions of normal areas while only a few sentences are about the abnormalities. This observation can explain why Ours-No-Attention can achieve very high scores to a certain degree: it can just generate the normalities to easily obtain a high score on n-gram bases evaluation systems.

As we dive into the content of the generated captions, it is surprising that different sentences have different topics. The first sentence is usually a high level description of the image, while each of the following sentences is associated with one area of the image (e.g. “lung”, “heart”, etc.). Besides, it is worth noting that Soft Attention and Ours-No-Attention models detect only a few abnormalities of the images, and the abnormalities detected by them are incorrect. In contrast, Ours-CoAttention model is able to correctly describe the many real abnormalities in the images (top three images). The comparison demonstrates that hierarchical LSTM is better for paragraph generation and model with co-attention is better at generating topics than model with no-attention.

For the third image, Ours-CoAttention model successfully detects the area (“right lower lobe”) which is abnormal (“eventration”), however, it fails to precisely describe this abnormality. In addition, the model also finds abnormalities about “interstitial opacities” and “atherosclerotic calcification”, which are not considered as real abnormality

Method	Normality	Abnormality	Total
Soft Attention	0.510	0.490	1.0
Ours-no-Attention	0.753	0.247	1.0
Ours-CoAttention	0.471	0.529	1.0
Ground Truth	0.385	0.615	1.0

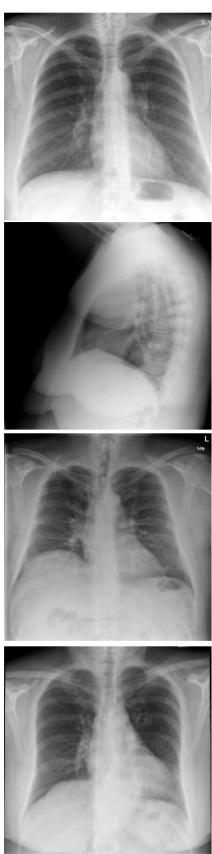
Table 2. Portion of sentences which describes the normalities and abnormalities in the image.

by human experts. The potential reason of mis-description might be that this x-ray image is darker (compared with the above images), and our model might be very sensitive to this change. The image at the bottom is a failure case of Ours-CoAttention. Even though the model makes the wrong judgment about the major abnormalities in the image, it does find some unusual regions: “lateral lucency” and “left lower lobe”. Additionally, it is also surprising to find that the model tries to reason about the findings by using “this may indicate”.

To better understand models’ ability of detecting real or potential abnormalities, we present the portion of sentences which describe the normalities and abnormalities in Table 2. Note that we consider sentences which contain “no” or “normal” or “clear” or “stable” as sentences describe the normalities. It is clear that Ours-CoAttention best approximates the distribution over normality and abnormality of the ground truth.

4.5.2 Co-Attention Learning

Figure 4 presents the visualization of co-attention. First property shown by Figure 4 is that sentence LSTM can generate different topics at different time steps since the model focuses on different regions of images and different tags for different sentences. The next finding is that visual attention can guide our model to concentrate on relevant regions



Ground Truth	Ours-CoAttention	Ours-no-Attention	Soft Attention
No active disease. The heart and lungs have in the interval. <u>Both lungs are clear and expanded.</u> Heart and mediastinum normal.	No active disease. The heart and lungs have in the interval. <u>Lungs are clear and expanded.</u> Cardiomedastinal silhouette is within normal limits. No pleural effusion or pneumothorax is seen. No cavity or pneumothorax.	The lungs are clear bilaterally. The are grossly normal. No focal lung consolidation. No acute bony abnormality. <u>cm nodule within the right lower lobe on the lateral view.</u> No pneumothorax or pleural effusion. Cardio mediastinal silhouette is unremarkable. The heart is not enlarged. The lungs are clear. No acute bony abnormality .	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. Specifically no evidence of focal airspace consolidation pleural effusion or pneumothorax. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.
No evidence of active disease. The lungs are clear. There is no focal airspace consolidation. No pleural effusion or pneumothorax. Heart size and mediastinal contour are within normal limits. <u>There are multilevel degenerative changes of the spine.</u>	No acute cardiopulmonary findings. Heart size is not enlarged. No focal airspace consolidation suspicious pulmonary opacity large pleural effusion or pneumothorax. No focal areas of consolidation. <u>Degenerative changes of the spine.</u> This is moderate exam of the hydropneumothorax. Lungs are clear. There is no focal airspace consolidation pleural effusion or pneumothorax.	The lungs are clear bilaterally. The are grossly normal. No pleural effusion. The heart is normal in size and contour. The lungs are clear. There are no acute bony findings.	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. There is no focal air space opacity to suggest a pneumonia.
No acute cardiopulmonary abnormality. Normal heart size mediastinal contours. Eventration of the right hemidiaphragm. No focal airspace consolidation. No pleural effusion or pneumothorax.	No acute cardiopulmonary abnormality. Stable appearance of the thoracic aorta. The right lateral lower lobe is noted in the <u>right lower right midlung.</u> No large pleural effusion or focal airspace disease. <u>Mild interstitial opacities, Atherosclerotic calcifications bony structures bilaterally.</u> There is no pleural effusion or pneumothorax developed in the right lower lobe.	The lungs are clear bilaterally. The are grossly normal. No acute bony abnormality. The lungs are otherwise clear. No acute osseous abnormality. No acute osseous abnormality. The heart and mediastinum are normal. There is no focal air space opacity.	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. There is no focal airspace consolidation. No pneumothorax or pleural effusion. Heart size and pulmonary vascularity appear within normal limits.
No acute cardiopulmonary abnormality. Heart size appears within normal limits. Pulmonary vasculature appears within normal limits. <u>Overlying the middle cardiac silhouette representing a hiatal hernia.</u> No focal consolidation pleural effusion or pneumothorax. No acute bony abnormality.	No active disease. The heart and lungs have in the interval. <u>Nipple and lateral lucency in the lungs suggestive of focal airspace disease.</u> The lungs are hyperexpanded consistent with <u>emphysema in the left lower lobe.</u> This is most at the upper lobes. This may indicate hypoventilated irregularities or effusions. The lungs are otherwise grossly clear. Resolution of by normal pleural effusion.	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. The are grossly normal. No focal airspace consolidation. No pneumothorax or pleural effusion. Heart size and pulmonary vascularity within normal limits. There is no pneumothorax or pleural effusion.	No acute cardiopulmonary abnormality. There is no focal airspace consolidation. No pneumothorax or pleural effusion. No acute bony abnormality. Heart size is normal .

Figure 3. Illustration of paragraph generated by Ours-CoAttention, Ours-no-Attention and Soft Attention models. The underlined sentences are the descriptions of detected abnormalities. The second image is a lateral x-ray image. Top two images are positive results, the third one is a partial failure case and the bottom one is failure case. These images are from test dataset.

of the image. For example, the third sentence of the first example is about “cardio”, and the visual attention concentrates on regions near the heart. Similar behavior can also be found for semantic attention: for the last sentence in first example, our model correctly concentrates on “degenerative change” which is the topic of the sentence. Besides, it is surprising that the content of first sentence in the second example contradicts the concentration of semantic attention. This is unlikely to happen for single attention mechanism. This contradiction implies that co-attention mechanism has certain fault tolerance, and thus co-attention might be more robust than single attention. Finally, the first sentence of last example presents a mis-description caused by incorrect semantic attention over tags. We believe incorrect attentions can be reduced by building a better tag prediction module.

4.6. Tag prediction

A diagnostic report contains not only text descriptions, but also tags which highlights key concepts in the report. We treat tag prediction as a multi-label classification task. To evaluate the prediction performance of multi-task learn-

ing, we compare our model and vanilla VGG-19 [17] on recall@5, recall@10, recall@20. Note that the loss function of both models are softmax cross-entropy loss [7].

Dataset	Methods	R@5	R@10	R@20
IU X-Ray	VGG-19 [17]	0.643	0.719	0.793
	Ours-CoAttention	0.644	0.716	0.792
PEIR Gross	VGG-19 [17]	0.392	0.506	0.595
	Ours-CoAttention	0.398	0.494	0.596

Table 3. Tag prediction on IU X-Ray dataset (upper part), and PEIR Gross dataset (lower part). R denotes recall.

Results in Table 3 show that Ours-CoAttention and VGG-19 [17] network performs very similar for tag prediction. Even though no improvement has been made by multi-task learning, our model is an end-to-end model which avoids managing a complex sequential pipeline.

Figure 4 also provides some qualitative results of tag prediction. The results show that in addition to tags which are relevant to the images, the model also produces many irrelevant tags. Even though co-attention mechanism can filter out many noisy tags, the irrelevant tags can still mislead the

					
degenerative change; obstruction	normal; <u>degenerative</u> change; nodule; calci- fied granuloma; hyper expansion; granuloma- tous disease; granu- loma; pneumonia; scarring; sternotomy	normal; <u>degenerative</u> change; nodule; calci- fied granuloma; hyper expansion; granuloma- tous disease; <u>granu-</u> loma; pneumonia; scar- ring; sternotomy	normal; <u>degenerative</u> change; nodule; calci- fied granuloma; hyper expansion; granuloma- tous disease; <u>granu-</u> loma; pneumonia; scar- ring; sternotomy	normal; <u>degenerative</u> change; nodule; calci- fied granuloma; hyper expansion; granuloma- tous disease; granu- loma; pneumonia; scar- ring; sternotomy	normal; <u>degenerative</u> change; nodule; calci- fied granuloma; hyper expansion; granuloma- tous disease; granu- loma; pneumonia; scar- ring; sternotomy
No acute intrathoracic abnormality.	No bony abnormality.	The cardio mediastinal silhouette is within nor- mal limits for appear- ance.	No focal areas of pul- monary consolidation.	Breast motion.	There is an age indeter- minate deformity of a mid-thoracic vertebral body.

No acute cardiopulmonary finding. The heart size and cardiopulmonary silhouette is normal. There is no focal airspace opacity pleural effusion or pneumothorax. The obstruction are intact with mild degenerative change in the thoracic spine.

					
calcified granuloma; granulomas	normal; <u>cardiomegaly</u> , nodule; degenera- tive change; <u>granulomato-</u> tous disease; opacity; <u>calci-</u> <u>fied granuloma</u> ; ate- lectasis; lymph nodes; hiatal hernia	normal; <u>cardiomegaly</u> , nodule; <u>degenerative</u> change; <u>granulomato-</u> tous disease; opacity; <u>calci- fied granuloma</u> ; ate- lectasis; lymph nodes; hiatal hernia	normal; <u>cardiomegaly</u> , nodule; <u>degenerative</u> change; <u>granulomato-</u> tous disease; opacity; <u>calci-</u> <u>fied granuloma</u> ; ate- lectasis; lymph nodes; hiatal hernia	normal; <u>cardiomegaly</u> , nodule; <u>degenerative</u> change; <u>granulomato-</u> tous disease; opacity; <u>calci-</u> <u>fied granuloma</u> ; ate- lectasis; lymph nodes; hiatal hernia	normal; <u>cardiomegaly</u> , nodule; <u>degenerative</u> change; <u>granulomato-</u> tous disease; opacity; <u>calci-</u> <u>fied granuloma</u> ; ate- lectasis; lymph nodes; hiatal hernia
Clear lungs.	Lungs are clear .	Left lower lung vol- umes is clear.	The heart is normal size.	A there is no focal air- space disease.	There is mild blunting of cost phrenic.

No acute cardiopulmonary abnormality. The lungs are clear. There are calcified granulomas. Heart size is normal. No pneumothorax.

					
normal	normal; <u>calcified gran-</u> uloma; <u>granuloma-</u> tous disease; granu- loma; scarring; opac- ity; degenera- tive change; sternotomy; thoracic aorta; nodule	normal; <u>calcified gran-</u> uloma; granulomatous disease; granuloma; scarring; opacity; <u>de-</u> <u>generative change</u> ;	normal; <u>calcified gran-</u> uloma; granulomatous disease; granuloma; scarring; opacity; <u>de-</u> <u>generative change</u> ;	normal; <u>calcified gran-</u> uloma; granulomatous disease; granuloma; scarring; opacity; <u>de-</u> <u>generative change</u> ;	normal; <u>calcified gran-</u> uloma; granulomatous disease; granuloma; scarring; opacity; <u>de-</u> <u>generative change</u> ;
Right upper lobe infil- trate.	Lungs are clear .	Stable heart size and aortic contours.	No acute displaced rib fractures.	No focal airspace opac- ities or consolidation.	No visualized of pneu- mothorax.

No acute cardiopulmonary abnormality identified. The examination consists of frontal and lateral radiographs of the chest. The cardio mediastinal contours are within normal limits. Pulmonary vascularity is within normal limits. No focal consolidation pleural effusion or pneumothorax identified. The visualized osseous structures and upper abdomen are unremarkable.

Figure 4. Visualization of co-attention for three examples. Each example is comprised of four lines: (1) image and visual attentions; (2) ground truth tags and semantic attention on predicted tags; (3) the generated descriptions; (4) ground truth descriptions. For the semantic attention, three tags with highest attention scores are highlighted. The underlined tags are the tags appear in the ground truth.

model to produce many false alarms. We believe a better tag prediction module will help model to attend to correct tags and thus help to improve the quality of generated captions.

5. Conclusion

In this paper, we study how to automatically generate textual reports for medical images, with the goal to help medical professionals produce reports more accurately and efficiently. Our proposed methods address three major challenges: (1) how to generate multiple heterogeneous forms of information, (2) how to localize abnormal regions and produce accurate descriptions for them, (3) how to gener-

ate long texts that contain multiple sentences or even paragraphs. To cope with these challenges, we propose a multi-tasking learning framework which jointly predicts tags and generates descriptions. We introduce a co-attention mechanism that can simultaneously explore visual and semantic information to accurately localize and describe abnormal regions. We develop a hierarchical LSTM network that can more effectively capture long-range semantics and produce high quality long texts. On two medical datasets containing radiology and pathology images, we demonstrate the effectiveness of the proposed methods through quantitative and qualitative studies.

References

- [1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *ICLR*, 2015. 3
- [2] L. Delrue, R. Gosselin, B. Ilsen, A. Van Landeghem, J. de Mey, and P. Duyck. Difficulties in the interpretation of chest radiography. In *Comparative Interpretation of CT and Standard Radiography of the Chest*, pages 27–49. Springer, 2011. 1
- [3] 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. 5
- [4] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5
- [5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 5, 6
- [6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 2
- [7] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *ICLR*, 2014. 7
- [8] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2
- [9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 2
- [10] P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul. From medical image to automatic medical report generation. *IBM Journal of Research and Development*, 59(2/3):2–1, 2015. 2
- [11] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 4
- [12] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition gan for visual paragraph generation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2
- [13] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004. 5
- [14] 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*, pages 311–318. Association for Computational Linguistics, 2002. 5
- [15] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. How to construct deep recurrent neural networks. *ICLR*, 2014. 4
- [16] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2016. 2
- [17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 5, 7
- [18] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5
- [19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2, 5, 6
- [20] 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*, pages 2048–2057, 2015. 2, 3, 4, 5, 6
- [21] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016. 2, 5, 6
- [22] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436, 2017. 2