On the Automatic Generation of Medical Imaging Reports

Baoyu Jing[1,2]   Pengtao Xie[1]   Eric P. Xing[1]

[1]Petuum Inc.   [2]Carnegie Mellon University

Introduction

Motivation

Medical imaging is widely used in clinical practice for diagnosis and treatment. Report-writing can be error-prone for unexperienced physicians, and time-consuming and tedious for experienced physicians. To address these issues, we study the automatic generation of medical imaging reports.

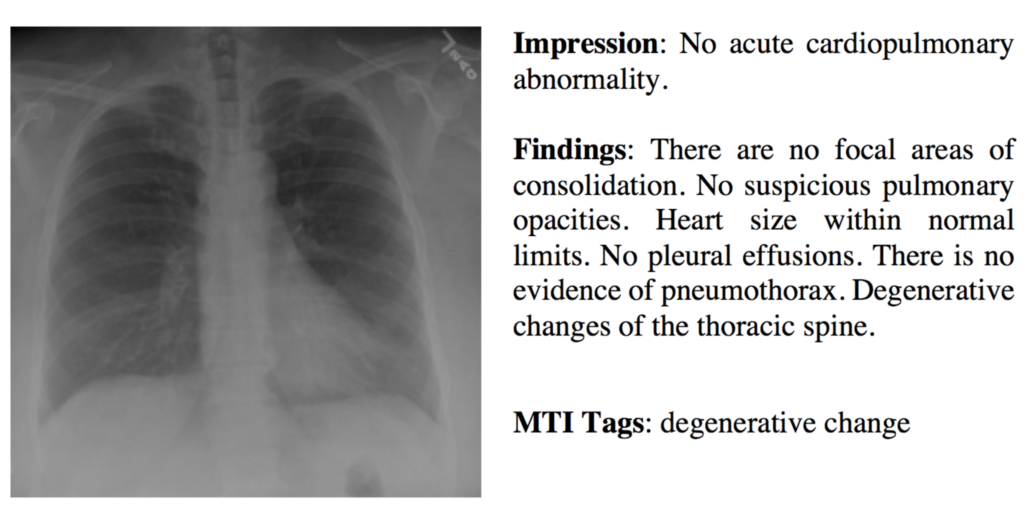

Figure 1: Example of Medical Report for a Chest X-ray Image

Challenges

(1) A report contains multiple heterogeneous forms of information, including *findings* and *tags*.
(2) Abnormal regions in medical images are difficult to identify.
(3) The reports are typically long and contain many sentences.

Contributions

(1) We build a multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs.
(2) We propose a co-attention mechanism to localize regions containing abnormalities and generate narrations for them.
(3) We develop a hierarchical LSTM model to generate long paragraphs.
(4) We perform extensive experiments to show the effectiveness of the proposed method.

Proposed Model

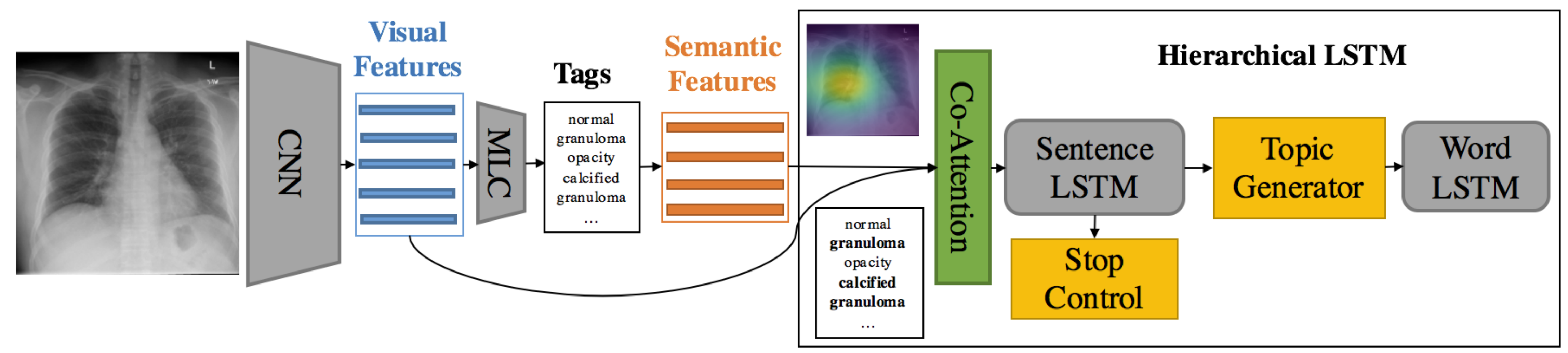

Figure 2: Overview of the Proposed Model

Encoding Process

Visual Information: (1) We use a CNN to learn visual features for different sub-regions of a given image. (2) These visual features are fed into a *multi-label classification* (MLC) network to predict relevant tags.
Semantic Information: (1) Each tag is represented by a word-embedding vector. (2) The word-embedding vectors of tags serve as the semantic features of this image.
Mix Visual & Semantic Information The visual and semantic features are fed into a *co-attention* model to generate a context vector that simultaneously captures the visual and semantic information.

Decoding Process

(1) **Sentence LSTM**: The context vector is input into the sentence LSTM, which produces topic vectors through *topic generator* and controls the termination through *stop control*. (2) **Word LSTM**: Given a topic vector, the word LSTM takes it as input and generates a sequence of words to form a sentence.

Datasets

Dataset	Description
X-Ray	A set of chest x-ray images paired with their corresponding diagnostic reports. The dataset contains 7,470 pairs of images and reports.
PEIR	The PEIR Gross dataset contains 7,442 image caption pairs from 21 different sub-categories.

Table 1: Dataset descriptions.

Tag Prediction

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

Table 2: Tag prediction on IU X-Ray and PEIR Gross dataset. R denotes recall.

Main Results

Dataset	Methods	B-1	B-2	B-3	B-4	Meteor	Rouge	Cider
X-Ray	CNN-RNN[1]	0.316	0.211	0.140	0.095	0.159	0.267	0.111
	LRCN[2]	0.369	0.229	0.149	0.099	0.155	0.278	0.190
	Soft ATT[3]	0.399	0.251	0.168	0.118	0.167	0.323	0.302
	ATT-RK[4]	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	CNN-RNN[1]	0.247	0.178	0.134	0.092	0.129	0.247	0.205
	LRCN[2]	0.261	0.184	0.136	0.088	0.135	0.254	0.203
	Soft ATT[3]	0.283	0.212	0.163	0.113	0.147	0.271	0.276
	ATT-RK[4]	0.274	0.201	0.154	0.104	0.141	0.264	0.279
	Ours-No-Attention	0.248	0.180	0.133	0.093	0.131	0.242	0.206
	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	0.280
	Ours-CoAttention	**0.300**	**0.218**	**0.165**	**0.113**	**0.149**	**0.279**	**0.329**

Table 3: 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.

Paragraph Generation

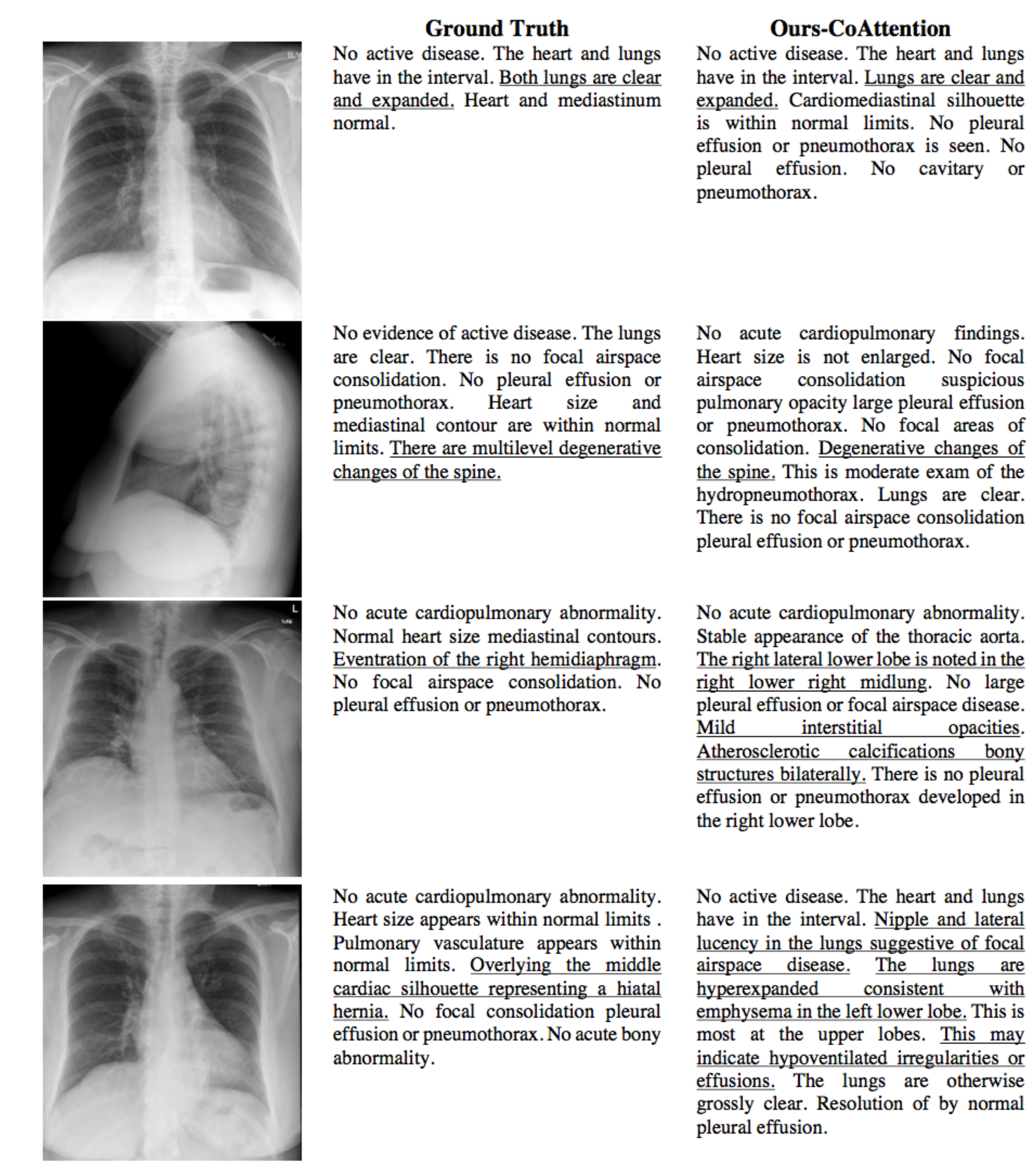

Figure 3: Examples of generated paragraphs.

Co-Attention Learning

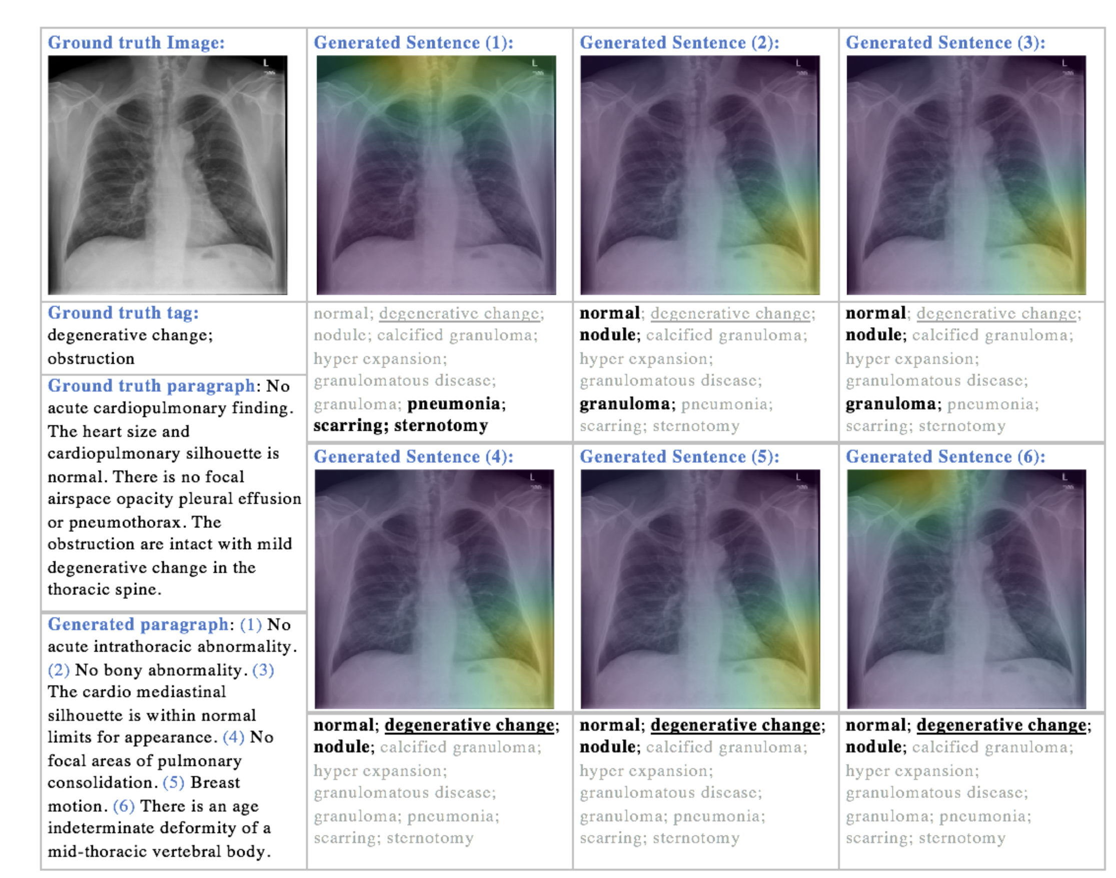

Figure 4: Co-Attention Learning.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *CVPR*

[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *ICML*

[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. *CVPR*