Show, Describe and Conclude:
On Exploiting the Structure Information of Chest X-Ray Reports

Baoyu Jing   Zeya Wang   Eric Xing

Petuum Inc., USA

{baoyu.jing, zeya.wang, eric.xing}@petuum.com

Abstract

Chest X-Ray (CXR) images are commonly used for clinical screening and diagnosis. Automatically writing reports for these images can considerably lighten the workload of radiologists for summarizing descriptive findings and conclusive impressions. The complex structures between and within sections of the reports pose a great challenge to the automatic report generation. Specifically, the section *Impression* is a diagnostic summarization over the section *Findings*; and the appearance of *normality* dominates each section over that of *abnormality*. Existing studies rarely explore and consider this fundamental structure information. In this work, we propose a novel framework which exploits the structure information between and within report sections for generating CXR imaging reports. First, we propose a two-stage strategy that explicitly models the relationship between *Findings* and *Impression*. Second, we design a novel cooperative multi-agent system that implicitly captures the imbalanced distribution between *abnormality* and *normality*. Experiments on two CXR report datasets show that our method achieves state-of-the-art performance in terms of various evaluation metrics. Our results expose that the proposed approach is able to generate high-quality medical reports through integrating the structure information.

1 Introduction

Chest X-Ray (CXR) image report generation aims to automatically generate detailed findings and diagnoses for given images, which has attracted growing attention in recent years (Wang et al., 2018a; Jing et al., 2018; Li et al., 2018). This technique can greatly reduce the workload of radiologists for interpreting CXR images and writing corresponding reports. In spite of the progress made in this area, it is still challenging for computers
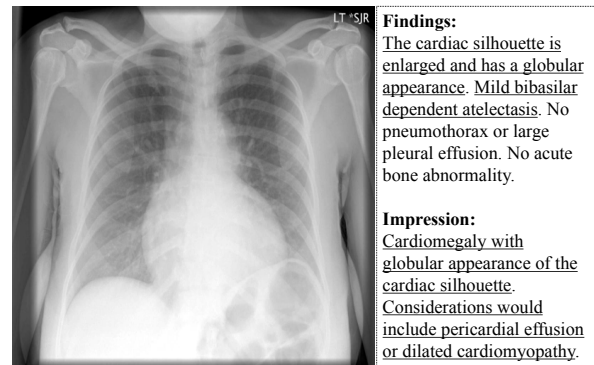

Figure 1: An example of chest X-ray image along with its report. In the report, the *Findings* section records detailed descriptions for normal and abnormal findings; the *Impression* section provides a diagnostic conclusion. The underlined sentence is an abnormal finding.

to accurately write reports. Besides the difficulties in detecting lesions from images, the complex structure of textual reports can prevent the success of automatic report generation. As shown in Figure 1, the report for a CXR image usually comprises two major sections: *Findings* and *Impression*. *Findings* section records detailed descriptions about normal and abnormal findings, such as lesions (e.g. increased lung marking). *Impression* section concludes diseases (e.g. pneumonia) from *Findings* and forms a diagnostic conclusion, consisting of abnormal and normal conclusions.

Existing methods (Wang et al., 2018a; Jing et al., 2018; Li et al., 2018) ignored the relationship between *Findings* and *Impression*, as well as the different distributions between normal and abnormal findings/conclusions. In addressing this problem, we present a novel framework for automatic report generation by exploiting the structure of the reports. Firstly, considering the fact that *Impression* is a summarization of *Findings*, we propose a two-stage modeling strategy given in Figure 3, where we borrow strength from im-

age captioning task and text summarization task for generating *Impression*. Secondly, we decompose the generation process of both *Findings* and *Impression* into the following recurrent sub-tasks: 1) examine an area in the image (or a sentence in *Findings*) and decide if an abnormality appears; 2) write detailed (normal or abnormal) descriptions for the examined area.

In order to model the above generation process, we propose a novel Co-operative Multi-Agent System (CMAS), which consists of three agents: Planner (PL), Abnormality Writer (AW) and Normality Writer (NW). Given an image, the system will run several loops until PL decides to stop the process. Within each loop, the agents co-operate with each other in the following fashion: 1) PL examines an area of the input image (or a sentence of *Findings*), and decides whether the examined area contains lesions. 2) Either AW or NW will generate a sentence for the area based on the order given by PL. To train the system, REINFORCE algorithm (Williams, 1992) is applied to optimize the reward (e.g. BLEU-4 (Papineni et al., 2002)). To the best of our knowledge, our work is the first effort to investigate the structure of CXR reports.

The major contributions of our work are summarized as follows. First, we propose a two-stage framework by exploiting the structure of the reports. Second, We propose a novel Co-operative Multi-Agent System (CMAS) for modeling the sentence generation process of each section. Third, we perform extensive quantitative experiments to evaluate the overall quality of the generated reports, as well as the model's ability for detecting medical abnormality terms. Finally, we perform substantial qualitative experiments to further understand the quality and properties of the generated reports.

2 Related Work

Visual Captioning The goal of visual captioning is to generate a textual description for a given image or video. For one-sentence caption generation, almost all deep learning methods (Mao et al., 2014; Vinyals et al., 2015; Donahue et al., 2015; Karpathy and Fei-Fei, 2015) were based on Convolutional Neural Network (CNN) - Recurrent Neural Network (RNN) architecture. Inspired by the attention mechanism in human brains, attention-based models, such as vi-

sual attention (Xu et al., 2015) and semantic attention (You et al., 2016), were proposed for improving the performances. Some other efforts have been made for building variants of the hierarchical Long-Short-Term-Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) to generate paragraphs (Krause et al., 2017; Yu et al., 2016; Liang et al., 2017). Recently, deep reinforcement learning has attracted growing attention in the field of visual captioning (Ren et al., 2017; Rennie et al., 2017; Liu et al., 2017; Wang et al., 2018b). Additionally, other tasks related to visual captioning, (e.g., dense captioning (Johnson et al., 2016), multi-task learning (Pasunuru and Bansal, 2017)) also attracted a lot of research attention.

Chest X-ray Image Report Generation Shin et al. (2016) first proposed a variant of CNN-RNN framework to predict tags (location and severity) of chest X-ray images. Wang et al. (2018a) proposed a joint framework for generating reference reports and performing disease classification at the same time. However, this method was based on a single-sentence generation model (Xu et al., 2015), and obtained low BLEU scores. Jing et al. (2018) proposed a hierarchical language model equipped with co-attention to better model the paragraphs, but it tended to produce normal findings. Despite Li et al. (2018) enhanced language diversity and model's ability in detecting abnormalities through a hybrid of template retrieval module and text generation module, manually designing templates is costly and they ignored the template's change over time.

Multi-Agent Reinforcement Learning The target of multi-agent reinforcement learning is to solve complex problems by integrating multiple agents that focus on different sub-tasks. In general, there are two types of multi-agent systems: independent and cooperative systems (Tan, 1993). Powered by the development of deep learning, deep multi-agent reinforcement learning has gained increasing popularity. Tampuu et al. (2017) extended Deep Q-Network (DQN) (Mnih et al., 2013) into a multi-agent DQN for Pong game; Foerster et al. (2016); Sukhbaatar et al. (2016) explored communication protocol among agents; Zhang et al. (2018) further studied fully decentralized multi-agent system. Despite these many attempts, the multi-agent system for long paragraph generation still remains unexplored.
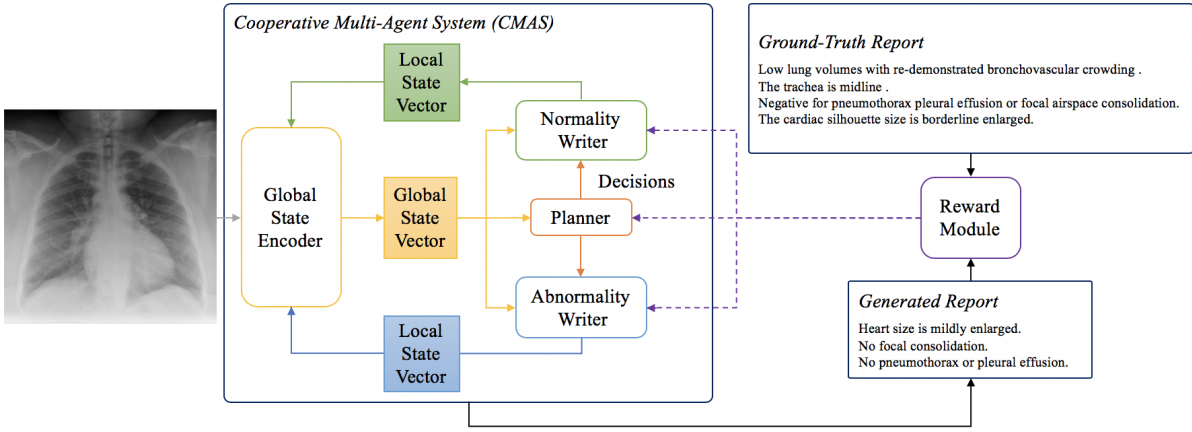
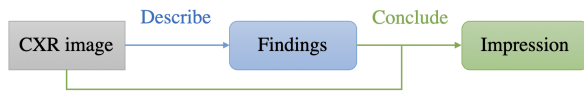Figure 2: Overview of the proposed Cooperative Multi-Agent System (CMAS).

Figure 3: Show, Describe and Conclude.

3 Overall Framework

As shown in Figure 3, the proposed framework is comprised of two modules: Findings and Impression. Given a CXR image, the Findings module will examine different areas of the image and generate descriptions for them. When findings are generated, the Impression module will give a conclusion based on findings and the input CXR image. The proposed two-stage framework explicitly models the fact that *Impression* is a conclusive summarization of *Findings*.

Within each module, we propose a Co-operative Multi-Agent System (CMAS) (see Section 4) to model the text generation process for each section.

4 Co-operative Multi-Agent System

4.1 Overview

The proposed Co-operative Multi-Agent System (CMAS) consists of three agents: Planner (PL), Normality Writer (NW) and Abnormality Writer (AW). These agents work cooperatively to generate findings or impressions for given chest X-ray images. PL is responsible for determining whether an examined area contains abnormality, while NW and AW are responsible for describing normality or abnormality in detail (Figure 2).

The generation process consists of several loops, and each loop contains a sequence of ac-

tions taken by the agents. In the n-th loop, the writers first share their local states $LS_{n-1,T} = \{w_{n-1,t}\}_{t=1}^{T}$ (actions taken in the previous loop) to form a shared global state $GS_n = (I, \{s_i\}_{i=1}^{n-1})$, where I is the input image, s_i is the i-th generated sentence, and $w_{i,t}$ is the t-th word in the i-th sentence of length T. Based on the global state GS_n, PL decides whether to stop the generation process or to choose a writer (NW or AW) to produce the next sentence s_n. If a writer is selected, then it will refresh its memory by GS_n and generate a sequence of words $\{w_{n,t}\}_{t=1}^{T}$ based on the sequence of local state $LS_{n,t} = \{w_{n,1}, \cdots, w_{n,t-1}\}$.

Once the generation process is terminated, the reward module will compute a reward by comparing the generated report with the ground-truth report. Given the reward, the whole system is trained via REINFORCE algorithm (Williams, 1992).

4.2 Policy Network

4.2.1 Global State Encoder

During the generation process, each agent will make decisions based on the global state GS_n. Since GS_n contains a list of sentences $\{s_i\}_{i=1}^{n-1}$, a common practice is to build a hierarchical LSTM as Global State Encoder (GSE) for encoding it. Equipping such an encoder with an excessive number of parameters for each agent in CMAS would be computation-consuming. We address this problem in two steps. First, we tie weights of GSE across the three agents. Second, instead of encoding previous sentences from scratch, GSE dynamically encodes GS_n based on GS_{n-1}. Specifically, we propose a single layer LSTM with soft-attention (Xu et al., 2015) as GSE. It takes a multi-modal context vector $\mathbf{ctx}_n \in \mathbb{R}^H$ as input,

which is obtained by jointly embedding sentence s_{n-1} and image I to a hidden space of dimension H, and then generates the global hidden state vector $\mathbf{gs}_n \in \mathbb{R}^H$ for the n-th loop by:

$$\mathbf{gs}_n = \text{LSTM}(\mathbf{gs}_{n-1}, \mathbf{ctx}_n) \qquad (1)$$

We adopt a visual attention module for producing context vector $\mathbf{ctx}_n$, given its capability of capturing the correlation between languages and images (Lu et al., 2017; Xu et al., 2015). The inputs to the attention module are visual feature vectors $\{\mathbf{v}_p\}_{p=1}^P \in \mathbb{R}^C$ and local state vector $\mathbf{ls}_{n-1}$ of sentence s_{n-1}. Here, $\{\mathbf{v}_p\}_{p=1}^P$ are extracted from an intermediate layer of a CNN, C and p are the number of channels and the position index of $\mathbf{v}_p$. $\mathbf{ls}_{n-1}$ is the final hidden state of a writer (defined in section 4.2.3). Formally, the context vector $\mathbf{ctx}_n$ is computed by the following equations:

$$\mathbf{h}_p = \tanh(\mathbf{W}_h[\mathbf{ls}_{n-1}; \mathbf{gs}_{n-1}]) \qquad (2)$$

$$\alpha_p = \frac{\exp(\mathbf{W}_{att}\mathbf{h}_p)}{\sum_{q=1}^P \exp(\mathbf{W}_{att}\mathbf{h}_q)} \qquad (3)$$

$$\mathbf{v}_{att} = \sum_{p=1}^P \alpha_p \mathbf{v}_p \qquad (4)$$

$$\mathbf{ctx}_n = \tanh(\mathbf{W}_{ctx}[\mathbf{v}_{att}; \mathbf{ls}_{n-1}]) \qquad (5)$$

where $\mathbf{W}_h$, $\mathbf{W}_{att}$ and $\mathbf{W}_{ctx}$ are parameter matrices; $\{\alpha_p\}_{p=1}^P$ are weights for visual features; and $[;]$ denotes concatenation operation.

At the beginning of the generation process, the global state is $GS_1 = (I)$. Let $\bar{\mathbf{v}} = \frac{1}{P}\sum_{i=1}^P \mathbf{v}_i$, the initial global state $\mathbf{gs}_0$ and cell state $\mathbf{c}_0$ are computed by two single-layer neural networks:

$$\mathbf{gs}_0 = \tanh(\mathbf{W}_{gs}\bar{\mathbf{v}}) \qquad (6)$$
$$\mathbf{c}_0 = \tanh(\mathbf{W}_c\bar{\mathbf{v}}) \qquad (7)$$

where $\mathbf{W}_{gs}$ and $\mathbf{W}_c$ are parameter matrices.

4.2.2 Planner

After examining an area, Planner (PL) determines: 1) whether to terminate the generation process; 2) which writer should generate the next sentence. Specifically, besides the shared Global State Encoder (GSE), the rest part of PL is modeled by a two-layer feed-forward network:

$$\mathbf{h}_n = \tanh(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{gs}_n)) \qquad (8)$$
$$idx_n = \arg\max(\text{softmax}(\mathbf{W}_3 \mathbf{h}_n)) \qquad (9)$$

where W_1, W_2, and W_3 are parameter matrices; $idx_n \in \{0, 1, 2\}$ denotes the indicator, where 0 is for STOP, 1 for NW and 2 for AW. Namely, if $idx_n = 0$, the system will be terminated; else, NW ($idx_n = 1$) or AW ($idx_n = 2$) will generate the next sentence s_n.

4.2.3 Writers

The number of normal sentences is usually 4-12 times to the number of abnormal sentences for each report. With this highly unbalanced distribution, using only one decoder to model all of the sentences would make the generation of normal sentences dominant. To solve this problem, we design two writers, i.e., Normality Writer (NW) and Abnormality Writer (AW), to model normal and abnormal sentences. Practically, the architectures of NW and AW can be different. In our practice, we adopt a single-layer LSTM for both NW and AW given the principle of parsimony.

Given a global state vector $\mathbf{gs}_n$, CMAS first chooses a writer for generating a sentence based on idx_n. The chosen writer will re-initialize its memory by taking $\mathbf{gs}_n$ and a special token BOS (Begin of Sentence) as its first two inputs. The procedure for generating words is:

$$\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{W}_e \mathbf{y}_{w_{t-1}}) \qquad (10)$$
$$\mathbf{p}_t = \text{softmax}(\mathbf{W}_{out}\mathbf{h}_t) \qquad (11)$$
$$w_t = \arg\max(\mathbf{p}_t) \qquad (12)$$

where $\mathbf{y}_{w_{t-1}}$ is the one-hot encoding vector of word w_{t-1}; $\mathbf{h}_{t-1}, \mathbf{h}_t \in \mathbb{R}^H$ are hidden states of LSTM; $\mathbf{W}_e$ is the word embedding matrix and $\mathbf{W}_{out}$ is a parameter matrix. $\mathbf{p}_t$ gives the output probability score over the vocabulary.

Upon the completion of the procedure (either token EOS (End of Sentence) is produced or the maximum time step T is reached), the last hidden state of LSTM will be used as local state vector $\mathbf{ls}_n$, which will be fed into GSE for generating next global state vector GS_{n+1}.

4.3 Reward Module

We use BLEU-4 (Papineni et al., 2002) to design rewards for all agents in CMAS. A generated paragraph is a collection $(\mathbf{s}^{ab}, \mathbf{s}^{nr})$ of normal sentences $\mathbf{s}^{nr} = \{s_1^{nr}, \ldots, s_{N_{nr}}^{nr}\}$ and abnormal sentences $\mathbf{s}^{ab} = \{s_1^{ab}, \ldots, s_{N_{ab}}^{ab}\}$, where N_{ab} and N_{nr} are the number of abnormal sentences and the number of normal sentences, respectively. Similarly, the

ground truth paragraph corresponding to the generated paragraph $(\mathbf{s}^{ab}, \mathbf{s}^{nr})$ is $(\mathbf{s}^{*ab}, \mathbf{s}^{*nr})$.

We compute BLEU-4 scores separately for abnormal and normal sentences. For the first n generated abnormal and normal sentences, we have:

$$f(s_n^{ab}) = \text{BLEU}(\{s_1^{ab}, \cdots, s_n^{ab}\}, \mathbf{s}^{*ab}) \quad (13)$$

$$f(s_n^{nr}) = \text{BLEU}(\{s_1^{nr}, \cdots, s_n^{nr}\}, \mathbf{s}^{*nr}) \quad (14)$$

Then, the immediate reward for s_n (s_n^{ab} or s_n^{nr}) is $r(s_n) = f(s_n) - f(s_{n-1})$. Finally, the discounted reward for s_n is defined as:

$$R(s_n) = \sum_{i=0}^{\infty} \gamma^i r(s_{n+i}) \quad (15)$$

where $\gamma \in [0, 1]$ denotes discounted factor, and $r(s_1) = \text{BLEU}(\{s_1\}, \mathbf{s}^*)$.

4.4 Learning

4.4.1 Reinforcement Learning

Given an input image I, three agents (PL, NW and AW) in CMAS work simultaneously to generate a paragraph $\mathbf{s} = \{s_1, s_2, \ldots, s_N\}$ with the joint goal of maximizing the discounted reward $R(s_n)$ (Equation 15) for each sentence s_n.

The loss of a paragraph $\mathbf{s}$ is negative expected reward:

$$L(\theta) = -\mathbb{E}_{n, s_n \sim \pi_\theta}[R(s_n)] \quad (16)$$

where π_θ denotes the entire policy network of CMAS. Following the standard REINFORCE algorithm (Williams, 1992), the gradient for the expectation $\mathbb{E}_{n, s_n \sim \pi_\theta}[R(s_n)]$ in Equation 16 can be written as:

$$\nabla_\theta L(\theta) = \mathbb{E}_{n, s_n \sim \pi_\theta}[R(s_n) \nabla_\theta - \log \pi_\theta(s_n, idx_n)] \quad (17)$$

where $-\log \pi_\theta(s_n, idx_n)$ is joint negative log-likelihood of sentence s_n and its indicator idx_n, and it can be decomposed as:

$$-\log \pi_\theta(s_n, idx_n)$$
$$= \mathbb{1}_{\{idx_n = AW\}} L_{AW} + \mathbb{1}_{\{idx_n = NW\}} L_{NW} + L_{PL}$$
$$= -\mathbb{1}_{\{idx_n = AW\}} \sum_{t=1}^{T} \log p_{AW}(w_{n,t})$$
$$- \mathbb{1}_{\{idx_n = NW\}} \sum_{t=1}^{T} \log p_{NW}(w_{n,t})$$
$$- \log p_{PL}(idx_n) \quad (18)$$

where L_{AW}, L_{NW} and L_{PL} are negative log-likelihoods; p_{AW}, p_{NW} and p_{PL} are probabilities of taking an action; $\mathbb{1}$ denotes indicator function.

Therefore, Equation 17 can be re-written as:

$$\nabla_\theta L(\theta) = \mathbb{E}_{n, s_n \sim \pi_\theta}[R(s_n)(\mathbb{1}_{\{idx_n = AW\}} \nabla L_{AW} + \mathbb{1}_{\{idx_n = NW\}} \nabla L_{NW} + \nabla L_{PL})] \quad (19)$$

4.4.2 Imitation Learning

It is very hard to train agents using reinforcement learning from scratch, therefore a good initialization for policy network is usually required (Bahdanau et al., 2016; Silver et al., 2016; Wang et al., 2018b). We apply imitation learning with cross-entropy loss to pre-train the policy network. Formally, the cross-entropy loss is defined as:

$$L_{CE}(\theta) = -\lambda_{PL} \sum_{n=1}^{N} \{\log p_{PL}(idx_n^*)\}$$
$$-\lambda_{NW} \sum_{n=1}^{N} \{\mathbb{1}_{\{idx_n^* = NW\}} \sum_{t=1}^{T} \log p_{NW}(w_{n,t}^*)\}$$
$$-\lambda_{AW} \sum_{n=1}^{N} \{\mathbb{1}_{\{idx_n^* = AW\}} \sum_{t=1}^{T} \log p_{AW}(w_{n,t}^*)\} \quad (20)$$

where w^* and idx^* denote ground-truth word and indicator respectively; λ_{PL}, λ_{NW} and λ_{AW} are balancing coefficients among agents; N and T are the number of sentences and the number of words within a sentence, respectively.

4.5 CMAS for Impression

Different from the Findings module, the inputs of the Impression module not only contain images I but also the generated findings $\mathbf{f} = \{f_1, f_2, \ldots, f_{N_f}\}$, where N_f is the total number of sentences. Thus, for the Impression module, the n-th global state becomes $GS_n = (I, \mathbf{f}, \{s_i\}_{i=1}^{n-1})$. The rest part of CMAS for the Impression module is exactly the same as CMAS for the Findings module. To encode $\mathbf{f}$, we extend the definition of multi-modal context vector $\mathbf{ctx}_n$ (Equation 5) to:

$$\mathbf{ctx}_n = \tanh(\mathbf{W}_{ctx}[\mathbf{v}_{att}; \mathbf{f}_{att}; \mathbf{ls}_{n-1}]) \quad (21)$$

where $\mathbf{f}_{att}$ is the soft attention (Bahdanau et al., 2014; Xu et al., 2015) vector, which is obtained similar as $\mathbf{v}_{att}$ (Equation 3 and 4).

5 Experiments

5.1 Datasets

IU-Xray Indiana University Chest X-Ray Collection (Demner-Fushman et al., 2015) is a public dataset containing 3,955 fully de-identified radiology reports collected from the Indiana Network for Patient Care, each of which is associated with a frontal and/or lateral chest X-ray images, and there are 7,470 chest X-ray images in total. Each report is comprised of several sections: *Impression*, *Findings* and *Indication* etc. We preprocess the reports by tokenizing, converting tokens into lowercases and removing non-alpha tokens.

CX-CHR CX-CHR (Li et al., 2018) is a proprietary internal dataset, which is a Chinese chest X-ray report dataset collected from a professional medical examination institution. This dataset contains examination records for 35,500 unique patients, each of which consists of one or multiple chest X-ray images as well as a textual report written by professional radiologists. Each textual report has sections such as *Complain*, *Findings* and *Impression*. The textual reports are preprocessed through tokenizing by "jieba"[1], a Chinese text segmentation tool, and filtering rare tokens.

5.2 Experimental Setup

Abnormality Term Extraction Human experts helped manually design patterns for most frequent medical abnormality terms in the datasets. These patterns are used for labeling abnormality and normality of sentences, and also for evaluating models' ability to detect abnormality terms. The abnormality terms in *Findings* and *Impression* are different to some degree. This is because many abnormality terms in *Findings* are descriptions rather than specific disease names. For examples, "low lung volumes" and "thoracic degenerative" usually appear in *Findings* but not in *Impression*.

Evaluation Metrics We evaluate our proposed method and baseline methods on: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). The results based on these metrics are obtained by the standard image captioning evaluation tool[2]. We also calculate precision and average False Positive Rate (FPR) for abnormality detection in generated textual reports on both datasets.

---

[1]https://github.com/fxsjy/jieba.
[2]https://github.com/tylin/coco-caption

Implementation Details The dimensions of all hidden states in Abnormality Writer, Normality Writer, Planner and shared Global State Encoder are set to 512. The dimension of word embedding is also set as 512.

We adopt ResNet-50 (He et al., 2016) as image encoder, and visual features are extracted from its last convolutional layer, which yields a $7 \times 7 \times 2048$ feature map. The image encoder is pretrained on ImageNet (Deng et al., 2009)). For the IU-Xray dataset, the image encoder is fine-tuned on ChestX-ray14 dataset (Wang et al., 2017), since the IU-Xray dataset is too small. For the CX-CHR dataset, the image encoder is fine-tuned on its training set. The weights of the image encoder are then fixed for the rest of the training process.

During the imitation learning stage, the cross-entropy loss (Equation 20) is adopted for all of the agents, where λ_{PL}, λ_{AW} and λ_{NW} are set as 1.0. We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5×10^{-4} for both datasets. During the reinforcement learning stage, the gradients of weights are calculated based on Equation 19. We also adopt Adam optimizer for both datasets and the learning rate is fixed as 10^{-6}.

Comparison Methods For *Findings* section, we compare our proposed method with state-of-the-art methods for CXR imaging report generation: CoAtt (Jing et al., 2018) and HGRG-Agent (Li et al., 2018), as well as several state-of-the-art image captioning models: CNN-RNN (Vinyals et al., 2015), LRCN (Donahue et al., 2015), AdaAtt (Lu et al., 2017), Att2in (Rennie et al., 2017). In addition, we implement several ablated versions of the proposed CMAS to evaluate different components in it: $CMAS_W$ is a single agent system containing only one writer, but it is trained on both normal and abnormal findings. $CMAS_{NW,AW}$ is a simple concatenation of two single agent systems $CMAS_{NW}$ and $CMAS_{AW}$, which are respectively trained on only normal findings and only abnormal findings. Finally, we show CMAS's performances with imitation learning (CMAS-IL) and reinforcement learning (CMAS-RL).

For *Impression* section, we compare our method with Xu et al. (2015): $SoftAtt_{vision}$ and $SoftAtt_{text}$, which are trained with visual input only (no findings) and textual input only (no images). We also report CMAS trained only on visual and textual input: $CMAS_{text}$ and $CMAS_{vision}$. Finally, we also compare CMAS-IL with CMAS-RL.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
CX-CHR	CNN-RNN (Vinyals et al., 2015)	0.590	0.506	0.450	0.411	0.577	1.580
	LRCN (Donahue et al., 2015)	0.593	0.508	0.452	0.413	0.577	1.588
	AdaAtt (Lu et al., 2017)	0.588	0.503	0.446	0.409	0.575	1.568
	Att2in (Rennie et al., 2017)	0.587	0.503	0.446	0.408	0.576	1.566
	CoAtt (Jing et al., 2018)	0.651	0.568	0.521	0.469	0.602	2.532
	HGRG-Agent (Li et al., 2018)	0.673	0.587	0.530	0.486	0.612	2.895
	$CMAS_W$	0.659	0.585	0.534	0.497	0.627	2.564
	$CMAS_{NW,AW}$	0.657	0.579	0.522	0.479	0.585	1.532
	CMAS-IL	0.663	0.592	0.543	0.507	0.628	2.475
	CMAS-RL	**0.693**	**0.626**	**0.580**	**0.545**	**0.661**	**2.900**
IU-Xray	CNN-RNN (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	0.306	0.294
	LRCN (Donahue et al., 2015)	0.223	0.128	0.089	0.067	0.305	0.284
	AdaAtt (Lu et al., 2017)	0.220	0.127	0.089	0.068	0.308	0.295
	Att2in (Rennie et al., 2017)	0.224	0.129	0.089	0.068	0.308	0.297
	CoAtt (Jing et al., 2018)	0.455	0.288	0.205	0.154	0.369	0.277
	HGRG-Agent (Li et al., 2018)	0.438	0.298	0.208	0.151	0.322	**0.343**
	$CMAS_W$	0.440	0.292	0.204	0.147	0.365	0.252
	$CMAS_{NW,AW}$	0.451	0.286	0.199	0.146	0.366	0.269
	CMAS-IL	0.454	0.283	0.195	0.143	0.353	0.266
	CMAS-RL	**0.464**	**0.301**	**0.210**	**0.154**	**0.362**	0.275

Table 1: Main results for findings generation on the CX-CHR (upper) and IU-Xray (lower) datasets. BLEU-n denotes the BLEU score that uses up to n-grams.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
CX-CHR	$SoftAtt_{text}$ (Xu et al., 2015)	0.112	0.044	0.016	0.005	0.142	0.038
	$SoftAtt_{vision}$ (Xu et al., 2015)	0.408	0.300	0.247	0.208	0.466	0.932
	$CMAS_{text}$	0.182	0.141	0.127	0.119	0.356	2.162
	$CMAS_{vision}$	0.415	0.357	**0.323**	**0.296**	**0.511**	**3.124**
	CMAS-IL	0.426	0.360	0.322	0.290	0.504	3.080
	CMAS-RL	**0.428**	**0.361**	**0.323**	0.290	0.504	2.968
IU-Xray	$SoftAtt_{text}$ (Xu et al., 2015)	0.179	0.047	0.006	0.000	0.161	0.032
	$SoftAtt_{vision}$ (Xu et al., 2015)	0.224	0.103	0.045	0.022	0.210	0.046
	$CMAS_{text}$	0.316	0.235	0.187	0.148	**0.537**	**1.562**
	$CMAS_{vision}$	0.379	0.270	0.203	0.151	0.513	1.401
	CMAS-IL	0.399	0.285	0.214	0.158	0.517	1.407
	CMAS-RL	**0.401**	**0.290**	**0.220**	**0.166**	0.521	1.457

Table 2: Main results for impression generation on the CX-CHR (upper) and IU-Xray (lower) datasets. BLEU-n denotes the BLEU score that uses up to n-grams.

5.3 Main Results

Comparison to State-of-the-art Table 1 shows results on the automatic metrics for the Findings module. On both datasets, CMAS outperforms all baseline methods on almost all metrics, which indicates its overall efficacy for generating reports that resemble those written by human experts. The methods can be divided into two different groups: single sentence models (CNN-RNN, LRCN, AdaAtt, Att2in) and hierarchical models (CoAtt, HGRG-Agent, CMAS). Hierarchical models consistently outperform single sentence models on both datasets, suggesting that the hierarchical models are better for modeling paragraphs. The leading performances of CMAS-IL and CMAS-RL over the rest of hierarchical models demonstrate the validity of our practice in exploiting the structure information within sections.

Ablation Study $CMAS_W$ has only one writer, which is trained on both normal and abnormal findings. Table 1 shows that $CMAS_W$ can achieve competitive performances to the state-of-the-art methods. $CMAS_{NW, AW}$ is a simple concatenation of two single agent models $CMAS_{NW}$ and $CMAS_{AW}$, where $CMAS_{NW}$ is trained only on normal findings and $CMAS_{AW}$ is trained only on abnormal findings. At test time, the final paragraph of $CMAS_{NW, AW}$ is simply a concatenation of normal and abnormal findings generated by $CMAS_{NW}$ and $CMAS_{AW}$ respectively. Surprisingly, $CMAS_{NW, AW}$ performs worse than $CMAS_W$ on the CX-CHR dataset. We believe the main reason is the missing communication protocol between the two agents, which could cause conflicts when they take actions independently. For example, for an image, NW might think "the

Dataset	CX-CHR				IU-Xray			
Methods	Li et al. (2018)	CMAS$_{NW,AW}$	CMAS-IL	CMAS-RL	Li et al. (2018)	CMAS$_{NW,AW}$	CMAS-IL	CMAS-RL
Precision	0.292	0.173	0.272	**0.309**	0.121	0.070	0.094	**0.128**
FPR	0.059	0.076	0.063	**0.051**	0.043	0.044	0.012	**0.007**

Table 3: Average precision and average False Positive Rate (FPR) for abnormality detection. (*Findings*)

Dataset	CX-CHR				IU-Xray			
Methods	CMAS$_{text}$	CMAS$_{vision}$	CMAS-IL	CMAS-RL	CMAS$_{text}$	CMAS$_{vision}$	CMAS-IL	CMAS-RL
Precision	0.067	0.171	0.184	**0.187**	0.054	0.160	0.162	**0.165**
FPR	**0.067**	0.142	0.170	0.168	**0.023**	0.024	0.024	0.024

Table 4: Average precision and average False Positive Rate (FPR) for abnormality detection. (*Impression*)

heart size is normal", while AW believes "the heart is enlarged". Such conflict would negatively affect their joint performances. As evidently shown in Table 1, CMAS-IL achieves higher scores than CMAS$_{NW, AW}$, directly proving the importance of communication between agents and thus the importance of PL. Finally, it can be observed from Table 1 that CMAS-RL consistently outperforms CMAS-IL on all metrics, which demonstrates the effectiveness of reinforcement learning.

Impression Module As shown in Table 2, CMAS$_{vision}$ and CMAS$_{text}$ have higher scores than SoftAtt$_{vision}$ and SoftAtt$_{text}$, indicating the effectiveness of CMAS. It can also be observed from Table 2 that images provide better information than text, since CMAS$_{vision}$ and SoftAtt$_{vision}$ exceed the scores of CMAS$_{text}$ and SoftAtt$_{text}$ to a large margin on most of the metrics. However, further comparison among CMAS-IL, CMAS$_{text}$ and CMAS$_{vision}$ shows that text information can help improve the model's performance to some degree.

5.4 Abnormality Detection

The automatic evaluation metrics (e.g. BLEU) are based on n-gram similarity between the generated sentences and the ground-truth sentences. A model can easily obtain high scores on these automatic evaluation metrics by generating normal findings (Jing et al., 2018). To better understand CMAS's ability in detecting abnormalities, we report its precision and average False Positive Rate (FPR) for abnormality term detection in Table 3 and Table 4. Table 3 shows that CMAS-RL obtains the highest precision and the lowest average FPR on both datasets, indicating the advantage of CMAS-RL for detecting abnormalities. Table 4 shows that CMAS-RL achieves the highest precision scores, but not the lowest FPR. However, FPR can be lowered by simply generating normal sentences, which is exactly the behavior of CMAS$_{text}$.

5.5 Qualitative Analysis

In this section, we evaluate the overall quality of generated reports through several examples. Figure 4 presents 5 reports generated by CMAS-RL and CMAS$_W$, where the top 4 images contain abnormalities and the bottom image is a normal case. It can be observed from the top 4 examples that the reports generated by CMAS-RL successfully detect the major abnormalities, such as "cardiomegaly", "low lung volumes" and "calcified granulomas". However, CMAS-RL might miss secondary abnormalities sometimes. For instance, in the third example, the "right lower lobe" is wrongly-written as "right upper lobe" by CMAS-RL. We find that both CMAS-RL and CMAS$_W$ are capable of producing accurate normal findings since the generated reports highly resemble those written by radiologists (as shown in the last example in Figure 4). Additionally, CMAS$_W$ tends to produce normal findings, which results from the overwhelming normal findings in the dataset.

5.6 Template Learning

Radiologists tend to use reference templates when writing reports, especially for normal findings. Manually designing a template database can be costly and time-consuming. By comparing the most frequently generated sentences by CMAS with the most used template sentences in the ground-truth reports, we show that the Normality Writer (NW) in the proposed CMAS is capable of learning these templates automatically. Several most frequently used template sentences (Li et al., 2018) in the IU-Xray dataset are shown in Table 5. The top 10 template sentences generated by NW are presented in Table 6. In general, the templates sentences generated by NW are similar to those top templates in ground-truth reports.

Ground Truth — **CMAS$_W$** — **CMAS-RL**

Row 1:

Ground Truth: **cardiomegaly is noted** . no pleural effusions . no pneumothorax . **there is perihilar prominence and interstitial opacification** .

CMAS$_W$: the heart size and mediastinal contours are within normal limits . the lungs are clear . no focal consolidation pleural effusion or pneumothorax . no pneumothorax . no acute bony abnormality .

CMAS-RL: **the heart is large** . no focal consolidation . no pneumothorax or pleural effusion . cardiomediastinal silhouette is unremarkable . and soft tissues are unremarkable .

Row 2:

Ground Truth: the trachea is midline . **the heart is large unchanged from prior exam** . **slightly widened mediastinum secondary to cardiomegaly and a tortuous aorta is accentuated by ap portable** . **there are low lung volumes causing bibasilar atelectasis and bronchovascular crowding** . the lungs do not demonstrate focal infiltrate or effusion . there is no pneumothorax . the visualized bony structures reveal no acute abnormalities .

CMAS$_W$: the heart size is normal . the lungs are clear . no pleural effusion or pneumothorax . no pneumothorax . no acute bony abnormality . no acute bony abnormality.

CMAS-RL: the heart is normal in size . **there are low lung volumes with bronchovascular crowding** . the mediastinum is unremarkable . there is no pleural effusion or pneumothorax . there is no focal air space effusions to suggest a areas .

Row 3:

Ground Truth: the lungs and pleural spaces show no acute abnormality . **stable right lower lobe calcified granulomas** . **thin lingular scar unchanged** . heart size and pulmonary vascularity within normal limits . surgical clips are visualized in the right upper quadrant .

CMAS$_W$: the heart size and mediastinal contours are within normal limits . no focal airspace disease . no acute bony abnormality . no acute bony abnormality . no acute bony abnormality .

CMAS-RL: **there is a calcified granuloma in the right upper lobe** . no focal consolidation . no pleural effusion or pneumothorax . heart size is normal . no pneumothorax .

Row 4:

Ground Truth: **moderate cardiomegaly** . **mild bilateral costophrenic blunting and fissural thickening interstitial opacities greatest in the central lungs** and . **dense right lower lobe nodule and right hilar calcifications suggest a previous granulomatous process** .

CMAS$_W$: the heart size and pulmonary vascularity appear within normal limits . no focal airspace disease . no acute bony abnormality . no acute bony findings . no acute bony abnormality . the lungs are intact.

CMAS-RL: **heart size is mildly enlarged** . no focal consolidation . no pneumothorax or pleural effusion . heart size and pulmonary vascularity within normal limits . no acute bony abnormality .

Row 5:

Ground Truth: the heart is normal in size and contour . the lungs are clear without evidence of infiltrate . there is no pneumothorax or effusion .

CMAS$_W$: the heart size and contour . the lungs are clear . no focal airspace consolidation . no pneumothorax . no acute bony abnormality

CMAS-RL: the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no pleural effusion or pneumothorax .

Figure 4: Examples of findings generated by CMAS-RL and CMAS$_W$ on IU-Xray dataset, along with their corresponding CXR images and ground-truth reports. Highlighted sentences are abnormal findings.

The lungs are clear.
Lungs are clear.
The lung are clear bilaterally.
No pneumothorax or pleural effusion.
No pleural effusion or pneumothorax.
There is no pleural effusion or pneumothorax.
No evidence of focal consolidation, pneumothorax, or pleural effusion.
No focal consolidation, pneumothorax or large pleural effusion.
No focal consolidation, pleural effusion, or pneumothorax identified..

Table 5: Most commonly used templates in IU-Xray. Template sentences are clustered by their topics.

The lungs are clear.
The heart is normal in size.
Heart size is normal.
There is no acute bony abnormality.
There is no pleural effusion or pneumothorax.
There is no pneumothorax.
No pleural effusion or pneumothorax.
There is no focal air space effusion to suggest a areas.
No focal consolidation.
Trachea no evidence of focal consolidation pneumothorax or pneumothorax.

Table 6: Top 10 sentences generated by CMAS. The sentences are clustered by their topics.

6 Conclusion

In this paper, we proposed a novel framework for accurately generating chest X-ray imaging reports by exploiting the structure information in the reports. We explicitly modeled the between-section structure by a two-stage framework, and implicitly captured the within-section structure with a novel Co-operative Multi-Agent System (CMAS) comprising three agents: Planner (PL), Abnormality Writer (AW) and Normality Writer (NW). The entire system was trained with REINFORCE algorithm. Extensive quantitative and qualitative experiments demonstrated that the proposed CMAS not only could generate meaningful and fluent reports, but also could accurately describe the detected abnormalities.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *56th Annual Meeting of Computational Linguistics (ACL)*, pages 2577–2586.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3337–3345. IEEE.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. *arXiv preprint arXiv:1703.07022*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, page 3.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, page 2.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multitask video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1273–1283.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 290–298.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. 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.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pages 2244–2252.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395.

Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning (ICML)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018a. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018b. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4213–4222.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

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. In *International conference on machine learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. *Proceedings of the Tenth International Conference on Machine Learning (ICML)*.