arXiv:1805.08298v1 [cs.CV] 21 May 2018

Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation

Christy Y. Li, Xiaodan Liang, Zhiting Hu, Eric P. Xing
Carnegie Mellon University and Petuum Inc.
{yli3, xiaodan1, zhitingh, epxing}@cs.cmu.edu

Abstract

Generating long and coherent reports to describe medical images poses challenges to bridging visual patterns with informative human linguistic descriptions. We propose a novel Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) which reconciles traditional retrieval-based approaches populated with human prior knowledge, with modern learning-based approaches to achieve structured, robust, and diverse report generation. HRGR-Agent employs a hierarchical decision-making procedure. For each sentence, a high-level *retrieval policy module* chooses to either retrieve a template sentence from an off-the-shelf template database, or invoke a low-level *generation module* to generate a new sentence. HRGR-Agent is updated via reinforcement learning, guided by sentence-level and word-level rewards. Experiments show that our approach achieves the state-of-the-art results on two medical report datasets, generating well-balanced structured sentences with robust coverage of heterogeneous medical report contents. In addition, our model achieves the highest detection accuracy of medical terminologies, and improved human evaluation performance.

1 Introduction

Beyond the traditional visual captioning task [36, 24, 38, 35, 31, 15] that produces one single sentence, generating long and topic-coherent stories or reports to describe visual contents (images or videos) has recently attracted research interests [34, 30, 16], posed as a more challenging and realistic goal towards bridging visual patterns with human linguistic descriptions. Particularly, report generation has several challenges to be resolved: 1) the generated report is a long narrative consisting of multiple sentences or paragraphs, which must have a plausible logic and consistent topics; 2) there is a presumed content coverage and specific terminology/phrases, depending on the task at hand. For example, a sports game report should describe competing teams, wining points, and outstanding players [34]. (3) the content ordering is very crucial. For example, a sports game report usually talks about the competition results before describing teams and players in detail.

As one of most representative and practical report generation task, the desired medical image report generation [13] must satisfy more critical protocols and ensure the correctness of medical term usage. As shown in Figure 1, a medical report consists of a *findings* section describing medical observations in details of both normal and abnormal features, an *impression* or *conclusion* sentence indicating the most prominent medical observation or conclusion, and *comparison* and *indication* sections that list patient's peripheral information. Among these sections, the *findings* section posed as the most important component, ought to cover contents of various aspects such as heart size, lung opacity, bone structure; any abnormality appearing at lungs, aortic and hilum; and potential diseases such as effusion, pneumothorax and consolidation. And, in terms of content ordering, the narrative of *findings* section usually follows a presumptive order, e.g. heart size, mediastinum contour followed by lung opacity, remarkable abnormalities followed by mild or potential abnormalities.

Preprint. Work in progress.

Comparison:

Indication: 60-year-old male with seizure, ethanol abuse

Findings: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

Impression: Blunting of the right costophrenic sulcus could be secondary to a pleural effusion versus scarring.

Findings:

[R]: The heart size is normal. There is mild effusion. No acute bony abnormalities.

[G]: The heart size normal. No pleural effusion or pneumothorax. No acute bony abnormalities.

[HRGR-Agent]: The heart size and mediastinal contours are normal. There is blunting of costophrenic sulcus suggesting a small effusion. No bony abnormalities.

Figure 1: An example of medical image report generation. The middle column is a report written by radiologists for the chest x-ray image on the left column. The right column contains three reports generated by a retrieval-based system (R), a generation-based model (G) and our proposed model (HRGR-Agent) respectively. The retrieval-based model correctly detects effusion while the generative model fails to. Our HRGR-Agent detects effusion and also describes supporting evidence.

State-of-the-art captioning generation models [36, 8, 38, 29] tend to perform poorly on medical report generation with specific content requirements due to several reasons. First, medical reports are usually dominated by normal findings, that is, a small portion of majority sentences usually forms a template database. For these normal cases, a retrieval-based system (e.g. directly perform classification among a list of majority sentences given image features) can perform surprisingly well due to the low variance of language. For instance, in Figure 1, a retrieval-based system correctly detects effusion from a chest x-ray image, while a generative model that generates word-by-word given image features, fails to detect effusion. On the other hand, abnormal findings which are relatively rare and remarkably diverse, however, are of higher importance. Current text generation approaches [13] often fail to capture the diversity of such small portion of descriptions, and pure generation pipelines are biased towards generating plausible sentences that look natural by the language model but poor at finding visual groundings [14]. On the contrary, a desirable medical report usually has to not only describe normal and abnormal findings, but also support itself by visual evidences such as location and attributes of the detected findings appearing in the image.

Inspired by the fact that radiologists often follow templates for writing reports and modify them accordingly for each individual case [5, 11, 10], we propose a Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) which is the first attempt to incorporate human prior knowledge with learning-based generation for medical reports. HRGR-Agent employs a *retrieval policy module* to decide between automatically generating sentences by a *generation module* and retrieving specific sentences from the template database, and then sequentially generates multiple sentences via a hierarchical decision-making. The template database is built based on human prior knowledge collected from available medical reports. To enable effective and robust report generation, we jointly train the *retrieval policy module* and *generation module* via reinforcement learning (RL) [26] guided by sentence-level and word-level rewards, respectively. Figure 1 shows an example generated report by our HRGR-Agent which correctly describes "a small effusion" from the chest x-ray image, and successfully supports its finding by providing the appearance ("blunting") and location ("costophrenic sulcus") of the evidence.

Our main contribution is to bridge rule-based (retrieval) and learning-based generation via reinforcement learning, which can achieve plausible, correct and diverse medical report generation. Moreover, our HRGR-Agenet has several technical merits compared to existing retrieval-generation-based models: 1) our retrieval and generation modules are updated and benefit from each other via policy learning; 2) the retrieval actions are regarded as a part of the generation whose selection of templates directly influences the final generated result. [39] instead, uses retrieved templates as hidden states for the generative model; 3) the generation module is encouraged to learn diverse and complicated sentences while the retrieval policy module learns template-like sentences, driven by distinct word-level and sentence-level rewards, respectively. Other work such as [20] still enforces the generative model to predict template-like sentences.

We conduct extensive experiments on two medical image report dataset [7]. Our HRGR-Agent achieves the state-of-the-art performance on both datasets under three kinds of evaluation metrics: automatic metrics such as CIDEr[28], BLEU[21] and ROUGE[17], human evaluation, and detection accuracy of medical terminologies. Experiments show that the generated sentences by HRGR-Agent shares a descent balance between concise template sentences, and complicated and diverse sentences. Code will be made available soon.

2 Related Work

Visual Captioning and Report Generation. Visual captioning aims at generating a descriptive sentence for images or videos. State-of-the-art approaches use CNN-RNN architectures and attention mechanisms [23, 36, 9, 38, 35, 19, 2, 24]. The generated sequence is usually short, describing the most prominent visual event, and is primarily rewarded by language fluency in practice. Generating reports that are informative and have multiple sentences [34, 13] poses higher requirements on content selection, relation generation and content ordering. State-of-the-art methods on report generation [13] are still remarkably cloning expert behavior, and incapable of diversifying language and depicting rare but prominent findings. Our approach prevents from mimicking teacher behavior by sparing the burden of automatic generative model with a template selection and retrieval mechanism, which by design promotes language diversity and better content selection.

Templates Based Sequence Generation. Some of the recent approaches bridged generative language approaches and traditional template-based methods. However, state-of-the-art approaches either treat a retrieval mechanism as latent guidance [39], the impact of which to text generation is limited, or still encourage the generation network to mimic template-like sequences [20].

Reinforcement Learning in Sequence Generation. Recently, reinforcement learning (RL) has been receiving increased popularity in sequence generation task such as visual captioning [18, 24, 15, 31], text summarization [22, 6], and machine translation [3]. Traditional methods use cross entropy loss which is prone to exposure bias [23] and do not necessarily optimize evaluation metrics such as CIDEr [28], ROUGE [17], BLEU [21] and METEOR [4], while reinforcement learning can directly use the evaluation metrics as reward and update model parameters via gradient descent. There has been some recent efforts [31, 37] devoted in applying hierarchical reinforcement learning (HRL) in captioning where sequence generation is broken down into several sub-tasks each of which targets at a chunk of words [31]. However, HRL for long report generation is still under-explored.

3 Approach

Medical image report generation aims at generating a report consisting of a sequence of sentences $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ given a set of medical images $\mathbf{I} = \{I_j\}_{j=1}^K$ of a patient case. Each sentence comprises a sequence of words $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,N}), y_{i,j} \in \mathbb{V}$ where i is the index of sentences, j the index of words, and $\mathbb{V}$ the vocabulary of all output tokens. In order to generate long and topic-coherent reports, we formulate the decoding process in a hierarchical framework that first produces a sequence of hidden sentence topics, and then predicts words of each sentence conditioning on each topic.

It is observed that doctors writing a report tend to follow certain patterns and reuse templates, while adjusting statements for each individual case when necessary. To mimic the procedure, we propose to combine retrieval and generation for automatic report generation. In particular, we first compile an off-the-shelf template database $\mathbb{T}$ that consists of a set of sentences that occur frequently in the training corpus. Such sentences typically describe general observations, and are often inserted into medical reports, e.g., "the heart size is normal" and "there is no pleural effusion or pneumothorax". (Table 1 provides more examples.)

As described in Figure 2, a set of images for each sample is first fed into a CNN to extract visual features which is then transformed into a context vector by an *image encoder*. Then a *sentence decoder* recurrently generates a sequence of hidden states $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M)$ which represent sentence topics. Given each topic state $\mathbf{q}_i$, a *retrieval policy module* decides to either automatically generate a new sentence by invoking a *generation module*, or retrieve an existing template from the template database. Both the retrieval policy module (that determines between automatic generation or template retrieval) and the generation module (that generates words) are making discrete decisions and be updated via the REINFORCE algorithm [33, 26]. We devise sentence-level and word-level rewards accordingly for the two modules, respectively.

3.1 Hybrid Retrieval-Generation Reinforced Agent

Image Encoder. Given a set of images $\{I_j\}_{j=1}^K$, we first extract their features $\{\mathbf{v}_j\}_{j=1}^K$ with a pretrained CNN, and then average $\{\mathbf{v}_j\}_{j=1}^K$ to obtain $\mathbf{v}$. The image encoder converts $\mathbf{v}$ into a context

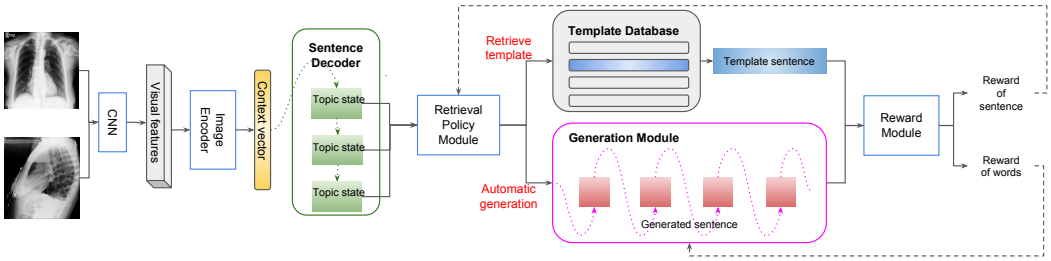


Figure 2: Hybrid Retrieval-Generation Reinforced Agent. Visual features are encoded by a CNN and image encoder, and fed to a sentence decoder to recurrently generate hidden topic states. A retrieval policy module decides for each topic state to either automatic generate a sentence, or retrieve a specific template from a template database. Dashed black lines indicate hierarchical policy learning.

vector $\mathbf{h}^v \in \mathbb{R}^D$ which is used as the visual input for all subsequent modules. Specifically, the image encoder is parameterized as a fully-connected layer, and the visual features are extracted from the last convolution layer of a DenseNet [12] or VGG-19 [25].

Sentence Decoder. Sentence decoder comprises stacked RNN layers which generates a sequence of topic states $\mathbf{q}$. We equip the stacked RNNs with attention mechanism to enhance text generation, inspired by [27, 36, 19]. Each stacked RNN first generates an attentive context vector $\mathbf{c}_i^s$, where i indicates time steps, given the image context vector $\mathbf{h}^v$ and previous hidden state $\mathbf{h}_{i-1}^s$. It then generates a hidden state $\mathbf{h}_i^s$ based on $\mathbf{c}_i^s$ and $\mathbf{h}_{i-1}^s$. The generated hidden state $\mathbf{h}_i^s$ is further projected into a topic space as $\mathbf{q}_i$ and a *stop control* probability $z_i \in [0, 1]$ through non-linear functions respectively. Formally, the sentence decoder can be written as:

$$\mathbf{c}_i^s = F_{\text{attn}}^s(\mathbf{h}^v, \mathbf{h}_{i-1}^s) \tag{1}$$
$$\mathbf{h}_i^s = F_{\text{RNN}}^s(\mathbf{c}_i^s, \mathbf{h}_{i-1}^s) \tag{2}$$
$$\mathbf{q}_i = \sigma(\mathbf{W}_q \mathbf{h}_i^s + \mathbf{b}_q) \tag{3}$$
$$z_i = \text{Sigmoid}(\mathbf{W}_z \mathbf{h}_i^s + \mathbf{b}_z), \tag{4}$$

where F_{attn}^s denotes a function of the attention mechanism [24], F_{RNN}^s denotes the non-linear functions of Stacked RNN, $\mathbf{W}_q$ and $\mathbf{b}_q$ are parameters which project hidden states into the topic space while $\mathbf{W}_z$ and $\mathbf{b}_z$ are parameters for stop control, and σ is a non-linear activation function. The stop control probability z_i greater than or equal to a predefined threshold (e.g. 0.5) indicates stopping generating topic states, and thus the hierarchical report generation process.

Retrieval Policy Module. Given each topic state $\mathbf{q}_i$, the retrieval policy module takes two steps. First, it predicts a probability distribution $\mathbf{u}_i \in R^{1+|\mathbb{T}|}$ over actions of generating a new sentence and retrieving from $|\mathbb{T}|$ candidate template sentences. Based on the prediction of the first step, it triggers different actions. If automatic generation obtains the highest probability, the generation module is activated to generate a sequence of words conditioned on current topic state (the second row on the right side of Figure 2). If a template in $\mathbb{T}$ obtains the highest probability, it is retrieved from the off-the-shelf template database and serves as the generation result of current sentence topic (the first row on the right side of Figure 2). We reserve 0 index to indicate the probability of selecting automatic generation and positive integers in $\{1, |\mathbb{T}|\}$ to index the probability of selecting templates in $\mathbb{T}$. The first step is parameterized as a fully-connected layer with Softmax activation:

$$\mathbf{u}_i = \text{Softmax}(\mathbf{W}_u \mathbf{q}_i + \mathbf{b}_u) \tag{5}$$
$$m_i = \text{argmax}(\mathbf{u}_i), \tag{6}$$

where $\mathbf{W}_u$ and $\mathbf{b}_u$ are network parameters, and the resulting m_i is the index of highest probability in $\mathbf{u}_i$.

Generation Module. Generation module generates a sequence of words conditioned on current topic state $\mathbf{q}_i$ and image context vector $\mathbf{h}^v$ for each sentence. It comprises RNNs which take environment parameters and previous hidden state $\mathbf{h}_{i,t-1}^g$ as input, and generate a new hidden state $\mathbf{h}_{i,t}^g$ which is further transformed to a probability distribution $\mathbf{a}_{i,t}$ over all words in $\mathbb{V}$, where t indicates t-th word. We define environment parameters as a concatenation of current topic state $\mathbf{q}_i$, context vector $\mathbf{c}_{i,t}^g$ encoded by following the same attention paradigm in sentence decoder, and embedding of previous

word $\mathbf{e}_{i,t-1}$. The procedure of generating each word is written as follows, which is an attentional decoding step:

$$\mathbf{c}^g_{i,t} = F^g_{\text{attn}}(\mathbf{h}^v, [\mathbf{e}_{i,t-1}; \mathbf{q}_i], \mathbf{h}^g_{i,t-1}) \tag{7}$$
$$\mathbf{h}^g_{i,t} = F^g_{\text{RNN}}([\mathbf{c}^g_{i,t}; \mathbf{e}_{i,t-1}; \mathbf{q}_i], \mathbf{h}^g_{i,t-1}) \tag{8}$$
$$\mathbf{a}_t = \text{Softmax}(\mathbf{W}_y \mathbf{h}^g_{i,t} + \mathbf{b}_y) \tag{9}$$
$$y_t = \text{argmax}(\mathbf{a}_t) \tag{10}$$
$$\mathbf{e}_{i,t} = \mathbf{W}_e \mathbb{O}(y_{i,t}), \tag{11}$$

where F^g_{attn} denotes the attention mechanism of generation module, F^g_{RNN} denotes non-linear functions of RNNs, $\mathbf{W}_y$ and $\mathbf{b}_y$ are parameters for generating word probability distribution, $y_{i,t}$ is index of the maximum probable word, $\mathbf{W}_e$ is a learnable word embedding matrix initialized uniformly, and $\mathbb{O}$ denotes one hot vector.

Reward Module. We use automatic metrics CIDEr [28] for computing rewards since recent work on image captioning [24] has shown that CIDEr performs better than many traditional automatic metrics such as BLEU [21], METEOR [4] and ROUGE [17]. We consider two kinds of reward functions: sentence-level reward and word-level reward. For the i-th generated sentencec $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,N})$ either from retrieval or generation outputs, we compute a delta CIDEr score at sentence level, which is $R_{sent}(\mathbf{y}_i) = f(\{\mathbf{y}_k\}^i_{k=1}, \text{gt}) - f(\{\mathbf{y}_k\}^{i-1}_{k=1}, \text{gt})$, where f denotes CIDEr evaluation, and gt denotes ground truth report. This assesses the advantages the generated sentence brings in to the existing sentences when evaluating the quality of the whole report. For a single word input, we use reward as delta CIDEr score which is $R_{word}(y_t) = f(\{y_k\}^t_{k=1}, \text{gt}^s) - f(\{y_k\}^{t-1}_{k=1}, \text{gt}^s)$ where gt^s denotes the ground truth sentence. The sentence-level and word-level rewards are used for computing discounted reward for retrieval policy module and generation module respectively.

3.2 Hierarchical Reinforcement Learning

Our objective is to maximize the reward of generated report $\mathbf{Y}$ compared to ground truth report $\mathbf{Y}^*$. Omitting the condition on image features for simplicity, the loss function can be written as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{z,m,y}[R(\mathbf{Y}, \mathbf{Y}^*)] \tag{12}$$
$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{z,m,y}\left[\nabla_\theta \log p(z, m, y) R(\mathbf{Y}, \mathbf{Y}^*)\right] \tag{13}$$
$$= -\mathbb{E}_{z,m,y}\left[\sum_{i=1} \mathbb{1}(z_i < \frac{1}{2} | z_{i-1}) \left(\nabla_{\theta_r} \mathcal{L}(\theta_r) + \mathbb{1}(m_i = 0 | m_{i-1}) \nabla_{\theta_g} \mathcal{L}(\theta_g)\right)\right], \tag{14}$$

where θ, θ_r ,and θ_g denote parameters of the whole network, *retrieval policy module*, and *generation module* respectively; $\mathbb{1}(\cdot)$ is binary indicator; z_i is the probability of topic stop control in Equation 4; m_i is the action chosen by *retrieval policy module* among automatic generation ($m_i = 0$) and all templates ($m_i \in [1, |\mathbb{T}|]$) in the template database. The loss of HRGR-Agent comes from two parts: *retrieval policy module* $\mathcal{L}(\theta_r)$ and *generation module* $\mathcal{L}(\theta_g)$ as defined below.

Policy Update for Retrieval Policy Module. We define the reward for retrieval policy module R^r at sentence level. The generated sentence or retrieved template sentence is used for computing the reward. The discounted sentence-level reward and its corresponding policy update according to REINFORCE algorithm [26] can be written as:

$$R^r(\mathbf{y}_i) = \sum_{j=0}^{\infty} \gamma^j R_{sent}(\mathbf{y}_{i+j}) \tag{15}$$
$$\mathcal{L}(\theta_r) = -\mathbb{E}_{m_i}[R^r(m_i, m^*_i)] \tag{16}$$
$$\nabla_{\theta_r} \mathcal{L}(\theta_r) = -\mathbb{E}_{m_i}[\nabla_{\theta_r} \log p(m_i | m_{i-1}) R^r(m_i, m^*_i)], \tag{17}$$

where γ is a discount factor; $\mathbf{y}_i$ is the i-th generated sequence; and θ_r represents parameters of retrieval policy module which are W_u and b_u in Equation 5 .

Policy Update for Generation Module. We define the word-level reward $R^g(y_t)$ for each word generated by generation module as discounted reward of all generated words after the considered

word. The discounted reward function and its policy update for generation module can be written as:

$$R^g(y_t) = \sum_{j=0}^{\infty} \gamma^j R_{word}(y_{t+j}) \tag{18}$$

$$\mathcal{L}(\theta_g) = -\mathbb{E}_{y_t}[R^g(\mathbf{y}_t, \mathbf{y}_t^*)] \tag{19}$$

$$\nabla_{\theta_g}\mathcal{L}(\theta_g) = -\mathbb{E}_{y_t}[\sum_{t=1} \nabla_{\theta_g} \log p(y_t|y_{t-1}) R^g(y_t, y_t^*)], \tag{20}$$

where γ is a discount factor, and θ_g represents the parameters of generation module such as W_y, b_y, W_e in Equation 9-11 and parameters of attention functions in Equation 7 and RNNs in Equation 8. Detailed policy update algorithm is provides in supplementary materials.

4 Experiments and Analysis

Datasets. We conduct experiments on two medical image report datasets. First, Indiana University Chest X-Ray Collection (IU X-Ray) [7] is a public dataset consists of 7,470 frontal and lateral-view chest x-ray images paired with their corresponding diagnostic reports. Each patient has 2 images and a report which includes impression, findings, comparison and indication sections. We treat the groundtruth report as a concatenation of impression and finding. We preprocess the reports by tokenizing, converting to lower-cases, and filtering tokens of frequency no less than 3. Following [13], the top 1000 most frequent tokens are selected the as vocabulary since they cover 99.0% word occurrences in the corpus. To fairly compare with all baselines [29, 8, 36, 38, 13], we extract visual features from the last convolutional layer of a VGG-19 model pretrained on classifying 572 unique tags that come with IU X-Ray dataset[13, 7], yielding $14 \times 14 \times 512$ feature maps.

CX-CHR is a private dataset of chest X-ray images with Chinese reports collected from a professional medical institution for health checking. The dataset consists of 35,500 patients. Each patient has one or multiple chest x-ray images in different views such as posteroanterior and lateral, and a corresponding Chinese report. We select patients with no more than 2 images and obtained 33,236 patient samples in total which covers over 93% of the dataset. To extract visual features, we pretrained a DenseNet with publically avaiable ChestX-ray8 dataset [32] on classification, and fine-tune it on CX-CHR dataset on 20 common thorax disease labels (see Supplementary Material for more details). Then we extract image features from the last convolutional layer, which yields $16 \times 16 \times 256$ feature maps. We preprocess the reports through tokenizing by Jieba [1] and filtering tokens of frequency no less than 3 as vocabulary, which results in 1282 unique tokens. We randomly split the dataset on patient-level into training, validation and testing by a raio of 7:1:2. We generate reports only in the findings section because it contains major and rich description of a report.

Template Database. We select sentences in the training set whose document frequencies (the number of occurrence of a sentence in training documents) are no less than a threshold as template candidates. We further group candidates that express the same meaning but have a little linguistic variations. For example, "no pleural effusion or pneumothorax" and "there is no pleural effusion or pneumonthorax" are grouped as one template. This results in 97 templates with greater than 500 document frequency for CX-CHR and 28 templates with greater than 100 document frequency for IU X-Ray. Upon retrieval, only the most frequent sentence of a template group will be retrieved for HRGR-Agent or any rule-based models that we compare with. Although this introduces minor but inevitable error in the generated results, our experiments show that the error is negligible compared to the advantages that a hybrid of retrieval-based and generation-based approaches brings in. Besides, separating templates of the same meaning into different categories diminishes the capability of *retrieval policy module* to predict the most suitable template for a given visual input, as multiple templates share the exact same meaning. Table 1 shows examples of templates for IU X-Ray dataset. More template examples are provided in supplementary materials.

Evaluation Metrics. We use three kinds of evaluation metrics: 1) automatic metrics including CIDEr [28], ROUGE [17], BLEU [21]; 2) medical abnormality terminology detection accuracy: we select 10 most frequent medical abnormality terminologies in medical reports and evaluate average detection accuracy and average false positive of compared models; 3) human evaluation: we randomly select 100 samples from testing set for each method and conduct surveys through Amazon Mechanical Turk. Each survey question gives a ground truth report, and ask candidate to choose among reports

Template	df(%)
No pneumothorax or pleural effusion. No pleural effusion or pneumothorax. There is no pleural effusion or pneumothorax.	18.36
The lungs are clear Lungs are clear. The lung are clear bilaterally.	23.60
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.	6.55
Cardiomediastin silhouett is within normal limit. The cardiomediastin silhouett is within normal limit. The cardiomediastin silhouett is within normal limit for size and contour.	5.12

Table 1: Examples of template database of IU X-Ray dataset. Each template is constructed by a group of sentences of the same meaning but slightly different linguistic variations. Top 3 most frequent sentences for a template are displayed in the first and third column. The second column shows document frequency (in percentage of training corpus) of each template.

generated by different models that matches with the ground truth report the best in terms of language fluency, content selection, and correctness of medical abnormal finding. A default choice is provided in case of no or both reports are preferred. We collect results from 20 participants and compute the average preference percentage for each model excluding default choices.

Training Details. We implement our model on PyTorch and train on a GeForce GTX TITAN GPU. We first train all models with cross entropy loss for 30 epochs with an initial learning rate of 5e-4, and then fine-tune the retrieval policy module and generation module of HRGR-Agent vis RL with a fixed learning rate 5e-5 for another 30 epochs. We use 512 as dimension of all hidden states and word embeddings, and batch size 16. We set the maxinum number of sentences of a report and maximum number of tokens in a sentence as 18 and 44 for CX-CHR and 7 and 15 for IU X-Ray. At testing, each generated report has average 7.2 and 4.8 sentences for CX-CHR and IU X-Ray dataset, respectively.

Baselines. For IU X-Ray dataset, we compare HRGR-Agent with 5 state-of-the-art image captioning models: CNN-RNN [29], LRCN [8], Soft ATT [36], ATT-RK [38] and CoAtt [13]. Visual features for all models are obtained from VGG-19 [25] for fair comparison. For CX-CHR dataset, we compare with 4 state-of-the-art methods: CNN-RNN [29], LRCN [8], AdaAtt [19] and Att2in [24]. Due to the relatively large size of CX-CHR, we conduct additional experiments on it to compare HRGR-Agent with its different variants by removing individual components (Retrieval, Generation, RL). We train a hierarchical generative model (*Generation*) without any template retrieval or RL fine-tuning, and our model without RL fine-tuning (HRG). To exam the quality of our pre-defined templates, we separately evaluate the *retrieval policy module* of HRGR-Agent by masking out the generation part and only use the retrieved templates as prediction (*Retrieval*). Note that *Retrieval* uses the same model as HRG-Agent whose training involves automatic generation of sentences, thus the results of which may be higher than a general retrieval-based system (e.g. directly perform classification among a list of majority sentences given image features).

4.1 Results and Analyses

Automatic Evaluation. Table 2 shows automatic evaluation comparison of state-of-the-art methods and our model variants. Most importantly, HRGR-Agent outperforms all baseline models (state-of-the-art methods that have no retrieval mechanism or hierarchical reinforcement learning) on both datasets by great margins, demonstrating its effectiveness and robustness. Particularly, on CX-CHR, HRGR-Agent increases CIDEr score by 0.73 compared to HRG, demonstrating that reinforcement fine-tuning is crucial to performance increase since it directly optimizes the evaluation metrics. Besides, *Retrieval* surpasses *Generation* by relatively large margins, showing that retrieval-based method is beneficial to generating structured reports, which leads to boosted performance of HRGR-Agent when combined with neural generation approaches (*generation module*).

Medical Terminology Accuracy. Table 3 shows evaluation results of average accuracy and average false positive of medical abnormality terminology detection. HGRG-Agent achieves the highest Acc.

Dataset	Model	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
CX-CHR	CNN-RNN [29]	1.580	0.590	0.506	0.450	0.411	0.577
	LRCN [8]	1.588	0.593	0.508	0.452	0.413	0.577
	AdaAtt [19]	1.568	0.588	0.503	0.446	0.409	0.575
	Att2in [24]	1.566	0.587	0.503	0.446	0.408	0.576
	Generation	0.361	0.3066	0.2159	0.1603	0.1205	0.3223
	Retrieval	2.565	0.5347	0.4754	0.4365	0.4094	0.5359
	HRG	2.800	0.6291	0.5470	0.4966	0.4626	0.5875
	HRGR-Agent	**3.530**	**0.6682**	**0.5849**	**0.5300**	**0.4855**	**0.6182**
IU X-Ray	CNN-RNN [29]	0.11	0.316	0.211	0.140	0.095	0.267
	LRCN [8]	0.190	0.369	0.229	0.149	0.099	0.278
	Soft ATT [36]	0.302	0.399	0.251	0.168	0.118	0.323
	ATT-RK[38]	0.155	0.369	0.226	0.151	0.108	0.323
	HRGR-Agent	**0.381**	**0.436**	**0.278**	**0.197**	**0.150**	**0.341**

Table 2: Automatic evaluation results on CX-CHR (upper part) and IU X-Ray Datasets (lower part). BLEU-n denotes BLEU score uses up to n-grams.

Dataset	CX-CHR			IU X-Ray		
Models	Retrieval	Generation	HRGR-Agent	CNN-RNN[29]	CoAtt[13]	HRGR-Agent
Acc. (%)	14.13	27.50	**29.19**	10.84	11.90	**12.13**
AFP	0.1333	0.0635	**0.059**	0.0237	0.082	**0.0428**
Hit (%)	–	23.42	**52.32**	–	28.00	**48.00**

Table 3: Average accuracy (Acc.) and average false positive (AFP) of medical abnormality terminology detection, and human evaluation (Hit). The higher Acc. and the lower AFP, the better.

and lowest AFP among all models, demonstrating that its robustness on detecting rare abnormal findings which are among the most important components of medical reports.

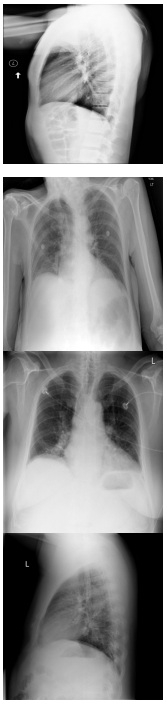
Retrieval vs. Generation. It's worth knowing that on CX-CHR, *Retrieval* achieves higher automatic evaluation scores (Table 2 the 7_{th} row) but lower medical term detection accuracy (Table 3 the 2_{nd} column) than *Generation*. Note that *Retrieval* evaluates *retrieval policy module* of HRGR-Agent by masking out the generation results of *generation module*. The result shows that simply composing templates that mostly describe normal medical findings can lead to high automatic evaluation scores since the majority reports describe normal cases. However, this kind retrieval-based approaches lack of the capability of detecting significant but rare abnormal findings. On the other hand, the high medical term detection accuracy of HRGR-Agent verifies that its *generation module* learns to describe abnormal findings. The win-win combination of *retrieval policy module* and *generation module* leads to state-of-the-art performance of HRGE-Agent, surpassing a generative model (*Generation*) that is purely trained without any retrieval mechanism.

Human Evaluation. Table 3 (last row) shows average human preference percentage of HRGR-Agent compared with *Generation* and CoAtt [13] on CX-CHR and IU X-Ray respectively, evaluated in terms of content coverage, specific terminology accuracy and language fluency. HRGR-Agent achieves much higher human preference than baseline models, showing that it is able to generate natural and plausible reports that are human preferable.

Qualitative Analysis. Figure 3 and 4 demonstrate qualitative results of HRGR-Agent and baseline models on both datasets. The reports of HRGR-Agent are generally longer than that of the baseline models, and share a well balance of templates and generated sentences. And, among the generated sentences, HRGR-Agent has higher rate of detecting abnormal findings.

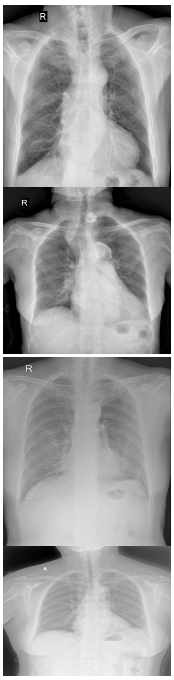
5 Conclusion

In this paper, we introduce a novel Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) to perform robust medical image report generation. Our approach is the first attempt to bridge human prior knowledge and generative neural network via reinforcement learning. Experiments show that HRGR-Agent does not only achieve state-of-the-art performance on two medical image report datasets, but also generates robust reports that has high accuracy on medical abnormal findings detection and best human preference.



Ground Truth	CoAtt [13]	HRGR-Agent
The cardiomediastinal silhouette is within normal limits. **Calcified right lower lobe granuloma**. No focal airspace consolidation. No visualized pneumothorax or large pleural effusion. No acute bony abnormalities.	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.	*The cardiomediastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits.* There is **right middle lobe airspace disease**, may reflect **granuloma** or pneumonia. *No pleural effusion. No pneumothorax. No acute bony abnormalities.*
Exam limited by patient rotation. Mild rightward **deviation of the trachea**. Stable **cardiomegaly**. Unfolding of the thoracic aorta. Persistent right **pleural effusion** with adjacent **atelectasis**. **Low lung volumes**. No focal airspace consolidation. There is severe **degenerative changes of the right shoulder**.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax. No acute bony abnormality.	**The heart is enlarged**. Possible **cardiomegaly**. There is pulmonary vascular congestion with diffusely increased interstitial and mild patchy airspace opacities. Suspicious **pleural effusion**. *There is no pneumothorax. There are no acute bony findings.*
Frontal and lateral views of the chest with overlying external cardiac monitor leads show **reduced lung volumes** with bronchovascular crowding of basilar **atelectasis**. No definite focal airspace consolidation or pleural effusion. **The cardiac silhouette appears mildly enlarged**.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax. no acute bony abnormality.	**The heart is mildly enlarged. The aorta is atherosclerotic and ectatic.** Chronic parenchymal changes are noted with mild scarring and/or subsegmental **atelectasis in the right lung base**. *No focal consolidation or significant pleural effusion identified.* **Costophrenic UNK are blunted**.
Apparent cardiomegaly partially accentuated by **low lung volumes**. No focal consolidation, pneumothorax or large pleural effusion. Right base **calcified granuloma**. Stable right infrahilar nodular density (lateral view). Negative for acute bone abnormality.	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.	*The heart size and pulmonary vascularity appear within normal limits.* **Low lung volumes**. Suspicious **calcified granuloma**. *No pleural effusion or pneumothorax. No acute bony abnormality.*

Figure 3: Examples of ground truth report and generated reports by CoAtt [13] and HRGR-Agent. Highlighted phrases are medical abnormality terms. Italicized text is retrieved from template database.



Ground Truth	Generation	Retrieval	HRGR-Agent
胸廓对称，诸肋完整。纵隔气管居中。**肺纹理增多、模糊**，肺内未见异常结节及肿块。**心影呈“靴形”，左缘第四弓增大。主动脉结内可见弧线形钙化，胸主动脉开展、迂曲**。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。膈下未见异常密度影。	胸廓对称，诸肋完整纵隔气管居中。双肺野清晰，纹理分布、走行自然。心影不大，**左心增大呈靴型改变**，心胸比约0，56，左室段圆钝。**主动脉增宽、迂曲扩**张，主动脉球突出。右侧肋膈角锐利。	胸廓对称，诸肋完整。纵隔气管居中。**双肺纹理增多、紊乱**，双肺未见明显实质性病灶。**主动脉结突出，钙化。**双侧肋膈角锐利。膈下未见异常密度影。	胸廓对称，诸肋完整。纵隔气管居中。**双肺纹理增多、紊乱**，双肺未见明显实质性病灶。**左心增大呈靴型改变。主动脉结突出，钙化。** 双侧肋膈角锐利。膈下未见异常密度影。
胸廓对称，诸肋完整，**上纵膈增宽。颈部见钙化灶**，双肺野清晰，纹理分布、走行自然，**右肺中野见结节。心影增大。主动脉见弧形钙化**。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。膈下未见异常密度影。	胸廓对称，诸肋完整。纵隔气管居中。右侧第5、8肋骨骨质断裂，断端错位性骨折征像错位。双肺野清晰，纹理分布、走行自然。心影不大。双侧膈肌位置、形态正常。	胸廓对称，诸肋完整。双肺野清晰，纹理分布、走行自然。**心影增大。主动脉迂曲、增宽**。双侧肋膈角锐利。膈下未见异常密度影。	胸廓对称，诸肋完整。**纵膈增宽**。双肺野清晰，纹理分布、走行自然。**右肺中野见结节密度影。心影增大。主动脉迂曲、增宽**。双侧肋膈角锐利。膈下未见异常密度影。
两胸廓对称。气管居中。**左侧下肺野见条索状密度增高影**，密度不均，边缘清晰，余两肺野内未见明显实质性病变，两肺野透亮度正常，**两肺纹理增多、模糊**。两肺门结构清晰。心脏大小、形态基本正常。双侧肋膈角清晰锐利。双侧膈面光滑。胸壁软组织及肋骨未见明显异常	胸廓对称，诸肋完整，**两肺纹理增多、模糊。** 右侧第8肋骨骨质断裂，断端错位性骨折征像，右侧第8后肋形态异常。纵隔气管居中，气管居中，纵膈无偏移，双肺内未见明确病灶影，边界清晰，无增粗、增多、变形，心影不大，形态正常，膈面光整。	胸廓对称，诸肋完整。纵隔气管居中。**两肺纹理增多、模糊**，肺内未见异常结节及肿块。心脏大小、形态基本正常。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。胸壁软组织及肋骨未见明显异常	胸廓对称，诸肋完整。纵隔气管居中。**两肺纹理增多、模糊，肺下野见密度增高条索影。**肺内未见异常结节及肿块。心脏大小、形态基本正常。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。胸壁软组织及肋骨未见明显异常
胸廓对称，诸肋完整。纵隔气管居中。**双肺纹理增多、增粗**，未见异常结节及肿块。心影不大。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。	胸廓对称，诸肋完整。纵膈居中。双肺野清晰，纹理分布、走行自然，肺内未见异常结节及肿块。心影不大，形态正常，膈面光整。双侧肋膈角锐利。	胸廓对称,诸肋完整。纵隔气管居中。心影不大。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。膈下未见异常密度影。	胸廓对称,诸肋完整。纵隔气管居中。**双肺纹理增多、紊乱**，未见异常结节及肿。心影不大。双侧膈肌位置、形态正常，膈面光整。双侧肋膈角锐利。膈下未见异常密度影。

Figure 4: Examples of ground truth report and generated reports by *Retrieval*, *Generation* and HRGR-Agent. The highlighted phrases are medical abnormality terminologies.

References

[1] "jieba" (chinese for "to stutter") chinese text segmentation: built to be the best python chinese word segmentation module. https://github.com/fxsjy/jieba. Accessed: 2018-05-01. 6

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*, 2018. 3

[3] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio. An actor-critic algorithm for sequence prediction. In *ICLR*, 2017. 3

[4] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005. 3, 5

[5] J. M. Bosmans, J. J. Weyler, A. M. De Schepper, and P. M. Parizel. The radiology report as seen by radiologists and referring clinicians: results of the cover and rover surveys. *Radiology*, 259(1):184–195, 2011. 2

[6] W. Chen, G. Li, S. Ren, S. Liu, Z. Zhang, M. Li, and M. Zhou. Generative bridging network in neural sequence prediction. In *NAACL*, 2018. 3

[7] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 2015. 2, 6

[8] 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 *CVPR*, 2015. 2, 6, 7, 8

[9] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *ICCV*, 2015. 3

[10] S. K. Goergen, F. J. Pool, T. J. Turner, J. E. Grimm, M. N. Appleyard, C. Crock, M. C. Fahey, M. F. Fay, N. J. Ferris, S. M. Liew, et al. Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *Journal of medical imaging and radiation oncology*, 57(1):1–7, 2013. 2

[11] Y. Hong and C. E. Kahn. Content analysis of reporting templates and free-text radiology reports. *Journal of digital imaging*, 26(5):843–849, 2013. 2

[12] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017. 4, 12

[13] B. Jing, P. Xie, and E. Xing. On the automatic generation of medical imaging reports. In *ACL*, 2018. 1, 2, 3, 6, 7, 8, 9

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[15] L. Li and B. Gong. End-to-end video captioning with multitask reinforcement learning. In *ICLR*, 2017. 1, 3

[16] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition gan for visual paragraph generation. In *ICCV*, 2017. 1

[17] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2013. 2, 3, 5, 6

[18] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, 2017. 3

[19] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 3, 4, 7, 8

[20] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018. 2, 3

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2, 3, 5, 6

[22] R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. In *ICLR*, 2018. 3

[23] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016. 3

[24] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1, 3, 4, 5, 7, 8

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 7

[26] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998. 2, 3, 5

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 4

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 3, 5, 6

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2, 6, 7, 8

[30] X. Wang, W. Chen, Y.-F. Wang, and W. Y. Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*, 2018. 1

[31] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, 2018. 1, 3

[32] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 6, 12

[33] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer, 1992. 3

[34] S. Wiseman, S. M. Shieber, and A. M. Rush. Challenges in data-to-document generation. In *ICCV*, 2017. 1, 3

[35] Z. Y. Y. Y. Y. Wu and R. S. W. W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In *NIPS*, 2016. 1, 3

[36] 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 *ICML*, 2015. 1, 2, 3, 4, 6, 7, 8

[37] D. Yarats and M. Lewis. Hierarchical text generation and planning for strategic dialogue. In *EMNLP*, 2017. 3

[38] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016. 1, 2, 3, 6, 7, 8

[39] S. L. Ziqiang Cao, Wenjie Li and F. Wei. Retrieve, rerank and rewrite: Soft template based neural summarization. In *ACL*, 2018. 2, 3

Appendices

A Policy Update Algorithm

Algorithm 1 describes policy update algorithm. If *retrieval policy module* predicts a template in $\mathbb{T}$, only *retrieval policy module* will be updated, by sentence-level reward. However, if *retrieval policy module* predicts automatic generation, *generation module* is also updated, by word-level reward. In practice, we alternatively train *retrieval policy module* and *generation module* while fixing another one. We use *Train-Generation* to indicate updating *generation module* in Algorithm 1.

Algorithm 1: Policy update procedure for HRGR-Agent

Data: Images $\{I_j\}$
Result: Generated report $\mathbf{Y} = (..., \mathbf{y}_i, ...)$

CNN extracts visual features;
image encoder extracts context vector;
for *time step* i **do**
  sentence decoder generates topic state $\mathbf{q}_i$ and z_i;
  if z_i *< 0.5* **then**
    retrieval policy module generates m_i;
    if m_i *== 0* **then**
      for *time step* t **do**
        generation module generates y_i;
      end
    else
      retrieve template indexed at m_i from template database;
    end
  end
end
for *reversed time step* i **do**
  if z_i *< 0.5* **then**
    if *Train-Generation* **then**
      if m_i *== 0* **then**
        for *reversed time step* t **do**
          reward module computes $R^g(y_i)$;
          update *generation module* by reward $R^g(y_i)$;
        end
      end
    else
      reward module computes $R^r(\mathbf{y}_i)$;
      update *retrieval policy module* by reward $R^r(\mathbf{y}_i)$;
    end
  end
end

B DenseNet Pretraining

We pretrain a DenseNet [12] with publically avaiable ChestX-ray8 dataset [32] on multi-label classification, and fine-tune it on CX-CHR dataset on 20 common thorax disease labels. ChestX-ray8 dataset [32] comprises 108,948 frontal-view X-ray images of 32,717 unique patients with each image labeled with occurrence of 14 common thorax diseases where labels were text-mined from the associated radiological reports using natural language processing techniques. We expand the 14 labels with 6 additional labels text-mined from CX-CHR dataset for fine-tuning. The additional 6 labels are: tortuous aortic sclerosis, bronchitis, calcification, tuberculosis, interstitial lung disease, and patchy consolidation.

We implement our model on PyTorch and train on a single GeForce GTX TITAN GPU. We add an additional transform layer after the last convolutional layer of DenseNet to convert features dimension to 256 for memory and computation efficiency, which yields $16 \times 16 \times 256$ feature maps. We use initial learning rate of 0.1 and multiply by 0.1 every 10 epochs. We train 30 epochs and select the best model by validation performance. The classification model achieves 78.04% AUC on test set.

We extract visual features from the last convolutional layer of the model for all experiments in our main paper.

C Template Database

Table 4 shows examples of template database of CX-CHR dataset. The template databases are designed by selecting the top most frequent sentences over a threshold in the training corpus and grouping sentences of the same meaning but slightly different language variation. The document frequency threshold for IU X-Ray and CX-CHR dataset is 100 and 500 respectively.

Template	df (%)	Template	df (%)
双侧肋膈角锐利 两侧肋膈角锐利 双肋膈角锐利	62.50	双侧胸廓对称 两侧胸廓对称 两胸廓对称	15.37
纵隔气管居中 气管、纵隔居中 气管纵隔居中 气管纵膈居中	61.30	心影大小、形态正常 心脏大小、形态正常 心脏形态、大小正常 心脏外形、大小正常	12.87
双侧膈面光整 双侧膈面光滑 两侧膈面光滑 两膈面光整	28.69	膈下未见异常密度影	31.28
		双肺纹理走形自然	2.59
		两肺纹理增重	2.44
		所见骨质无明显异常	1.83

Table 4: Examples of template database of CX-CHR dataset. Each template is constructed by a group of sentences of the same meaning but slightly different expressions. The second and third column display document frequency of individual sentence and template where all its sentences are included respectively. For a selected template at the retrieval step, only the first sentence is returned.