IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning

DAIBING HOU[1], ZIJIAN ZHAO[1], (Member, IEEE), YUYING LIU[1], FALIANG CHANG[1], AND SANYUAN HU [2]

[1]School of Control Science and Engineering, Shandong University, Jinan 250061, China
[2]Department of General surgery, First Affiliated Hospital of Shandong First Medical University, Jinan 250014, China

Corresponding author: Zijian Zhao (zhaozijian@sdu.edu.cn).

This work was supported by the National Key Research and Development Program of China under Grant 2019YFB1311300.

ABSTRACT An adversarial reinforced report-generation framework for chest x-ray images is proposed. Previous medical-report-generation models are mostly trained by minimizing the cross-entropy loss or further optimizing the common image-captioning metrics, such as CIDEr, ignoring diagnostic accuracy, which should be the first consideration in this area. Inspired by the generative adversarial network, an adversarial reinforcement learning approach is proposed for report generation of chest x-ray images considering both diagnostic accuracy and language fluency. Specifically, an accuracy discriminator (AD) and fluency discriminator (FD) are built that serve as the evaluators by which a report based on these two aspects is scored. The FD checks how likely a report originates from a human expert, while the AD determines how much a report covers the key chest observations. The weighted score is viewed as a "reward" used for training the report generator via reinforcement learning, which solves the problem that the gradient cannot be passed back to the generative model when the output is discrete. Simultaneously, these two discriminators are optimized by maximum-likelihood estimation for better assessment ability. Additionally, a multi-type medical concept fused encoder followed by a hierarchical decoder is adopted as the report generator. Experiments on two large radiograph datasets demonstrate that the proposed model outperforms all methods to which it is compared.

INDEX TERMS Medical report generation, encoder-decoder, adversarial training, reinforcement learning.

I. INTRODUCTION

Automatic radiology-report generation is a computer-aided diagnostic technology used for generating a free-text description of disease diagnosis or future treatment based on radiology images (such as chest x-rays). Compared with general disease diagnosis technology, it is closer to artificial intelligence (AI), for it can not only output a list of numbers corresponding to the probabilities of possible diseases but also "write" an easy-to-understand report with natural language. With this technology, patients can read the chest x-rays by themselves, and no longer have to queue up to consult doctors. Moreover, the workload of radiologists will be greatly lightened.

Chest x-rays are the most common type of radiology image, which produces images of the heart, lungs, airways, blood vessels, and bones of the spine and chest, and is used for diagnosis and treatment of chest diseases, such as pneumonia and pneumothorax. A chest x-ray report example is shown in Figure 1. Such a report includes two important parts: findings and impression. The former part describes in detailed the representations of different organs and regions and a determination of whether the patient has a certain or potential disease. The latter part is only the conclusion of the former part. Hence, the focus in this paper is on generation of the findings.

A similar study area is natural image captioning in computer vision and natural language processing because it has the same objective of mapping from images to text sequences. Hence, some common points exist between the two studies. First, encoder-decoder architecture is the basic architecture used to tackle these problems, in which the encoder, composed of a deep convolutional neural network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056175, IEEE Access

Findings:
There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with linear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

Impression:
Mild pulmonary edema.

Observations:
Cardiomegaly, Lung Opacity, Atelectasis.

MeSH Tags:
Pulmonary Atelectasis, Lung, Heart, Edema.

FIGURE 1. Example of chest x-ray images with corresponding report and post-extracted medical labels. Solid box shows the origin radiology report and dashed box the post-extracted medical labels used in the proposed approach.

(CNN), encodes images into a contextual vector, and the decoder, composed of long short-term memory (LSTM) [1], decodes the contextual vector into a word sequence step by step. Additionally, the entire model is usually trained by minimizing the cross-entropy loss or further retuned via reinforcement learning. However, an image-captioning approach cannot be directly applied to medical report generation for several significant reasons: (1) compared to natural images, chest x-rays involve complex and abstract medical concepts, e.g., “pulmonary atelectasis" and “cardiomegaly" shown in Figure 1, which is difficult for a plain encoder to capture; (2) a natural image caption mostly has one sentence, while a finding contains four, five, or even more sentences, and thus the basic decoder may struggle to learn such long-term dependencies; and (3) for medical report generation, one should make diagnostic accuracy the top priority, rather than blindly seeking high text-relevance scores (such as BLEU score).

Following [2]–[5], some improvements are made herein based on the original encoder-decoder approach. First, multi-type medical concepts are incorporated into the encoder. Detailed common chest observations and medical subject heading (MeSH) labels are adopted as two types of intermediate semantic information, which are predicted by a separate sub-network (called multi-label classiciation, or MLC) in the encoder. These predicted medical concepts will be embedded in the follow-up decoder along with the encoded image features. These two types of medical concepts have different semantic granularities, in which the observations cover generalized diseases, while the MeSH terms narrow their concepts to medical vocabulary. Second, a hierarchical LSTM is adopted as the decoder. The hierarchical decoder splits the decoding process into two stages: given encoded features, the sentence LSTM decodes topic vectors one by one, and then the word LSTM decodes a word sequence from a topic vector. Furthermore, the attention mechanism [6], [7] is also applied to the two decoding stages. More importantly, to achieve highly accurate and fluent medical reports, adversarial reinforcement learning (ARL) is introduced. Specifically, two additional discriminators are developed as the evaluator to provide an overall score considering both accuracy and fluency of a generated report. Different from the generative adversarial network (GAN) in image generation, the output of the report generator is a discrete sequence of words sampled from decision probabilities, and it blocks the transmission of the gradient from the discriminator to the generator [8], [9]. To solve this problem, reinforcement learning (RL) is introduced to optimize the generator. In adversarial training, the report generator is trained by RL viewing the overall score as the reward, and simultaneously the discriminators improve their judgment through maximum-likelihood estimation. The discriminators are the language fluency discriminator (FD) and diagnostic accuracy discriminator (AD), allowing for readability and accuracy, respectively. The FD checks how likely a report originates from a human expert, while the AD determines how much a report covers the ground-truth observations.

To the best of our knowledge, this is the first introduction of ARL to report generation for medical images. In short, the main contributions of the proposed framework are summarized as follows. (1) A novel ARL framework (Figure 2) is proposed in which the report generator is trained by RL with the rewards provided by discriminators, and the discriminators are updated via maximum-likelihood estimation in training iterations. (2) The proposed model is evaluated on two large chest-radiograph datasets with common captioning metrics and diagnostic accuracy, and it is found that the proposed model achieves the best performance against compared methods. (3) The medical concept prediction model (MLC branch) is embedded in the encoder, and, accordingly, the proposed model can not only generate a free-text report, but also predict observations compared to previous studies.

II. RELATED WORK

In this section, related research on natural image captioning, radiology-report generation, and adversarial training is reviewed.

A. NATURAL IMAGE CAPTIONING

The basic encoder-decoder architecture (also called CNN-RNN) for natural image captioning was first proposed by Vinyals et al. in 2015, overturning the previous template-based approach [10]. In this approach, the CNN-composed encoder maps an image into a context vector representation, and then the LSTM composed decoder unrolls and outputs the word distribution at every time step conditioned on the context vector. This model is trained by minimizing negative

log-likelihood or cross-entropy. Later, Xu et al. introduced the attention mechanism from machine translation to the decoder and achieved better performance [7]. All the above approaches are top-down, which start from an image and convert it into words, while in You et al.âĂŹs study another bottom-up approach was introduced that derives words describing various aspects of an image; in addition, semantic attention was also adopted [11]. For generating longer image captions, a hierarchical decoder was utilized to produce more detailed and coherent paragraph descriptions [12]. Prior models are mostly trained by minimizing negative log-likelihood or cross-entropy, but they suffer exposure bias and wrong-objective problems [13], among which the former means that the model is only exposed to the training data distribution instead of its own prediction, and the latter means cross-entropy loss trains the model to be adept at greedily predicting the next word at each time step without considering the entire sequence. Rennie et al. proposed a novel self-critical [14] training method that views the model as an agent and optimizes the CIDEr score by reinforcement learning. This work proves that RL is a very worthwhile recipe with which to boost the performance of captioning models.

B. RADIOLOGY-REPORT GENERATION

The achievements in natural image captioning have greatly promoted the development of medical report generation. The encoder-decoder architecture is also widely used in this area. To enforce the coherence between sentences, Xue et al. built an iterative decoder with visual attention [15], while Jing et al. introduced the hierarchical decoder and proposed a co-attention mechanism [2] that simultaneously attends to images and predicted tags. Additional works relevant to ours originate from Yuan et al. [5] and Liu et al. [3]. Yuan et al. extracted medical concepts as intermediate semantic features from images to enrich the decoder, which reduced the difficulty of direct mapping from medical images to reports. The medical concept prediction and visual feature encoding shared the same networks and only attended one type of concept feature into the decoder. In our work, a separate MLC model used for predicting multiple medical labels including observations and MeSH terms is built, and these predictions cannot only be fused into follow-up decoder, but also serve as one part of the final output for disease diagnosis. Liu et al. retuned their model by RL by designing two kinds of rewards: a natural language generation reward (NLGR) and a clinically coherent reward (CCR) that consider readability and accuracy, respectively. However, the reward design is empirically based and computationally complex, and cannot be improved during training. In the proposed approach, the reward is generated from learnable discriminators and can be updated in adversarial training.

C. ADVERSARIAL TRAINING

The GAN invented by Goodfellow et al. is one of unsupervised deep-learning systems [16]. Its general idea is that to train a generator (G) one selects another network as a discriminator (D), and the two neural networks compete with each other in the training process. G tries to generate a more realistic data distribution to deceive D, and D aims to estimate the probability of a sample come from real distribution rather than G. Owing to its significant effectiveness, adversarial training has become a general training methodology widely used in all kinds of generative models. Hundreds of GAN variants have been put forward for image generation in recent years, such as DCGAN [17] and StackGAN [18]. Different from image generation, sequence generation must go through a sampling operation, which is non-differentiable and cannot be optimized by gradient descent. Yu et al. proposed SeqGAN [9] and optimized the sequence generator using RL to overcome this problem, which is similar to the approach proposed in the present paper.

III. METHODOLOGY

A. OVERVIEW

In general, the proposed model can be divided into three components: encoder, decoder, and reward module, as shown in Figure 2. The encoder is comprised of two separate branches (CNN and MLC) that extract visual and semantic features for the decoder separately. The MLC branch, serving as a multi-label classification task, predicts common observations and other medical concepts for given images. These predicted medical labels are embedded in vectors and then inserted into the decoder. The decoder is developed by hierarchical LSTM [1], [12] with multi-level attention [5], [19]: the sentence LSTM generates topics vectors step by step and the word LSTM generates words one by one based on a topic vector. In adversarial training, the reward module composed of two discriminators generates a reward according to the quality of a generated report, which is then used to train the generator via RL. The decoder and reward module, as a pair of adversaries, are updated alternatively in training iterations. The details are elaborated in the following subsections.

B. ENCODER

Prior works [2], [5], [11], [20], [21] have proved that incorporating semantic information can improve the performance of natural image captioning and medical-image-report generation. These semantic labels are predicted commonly by multi-label classification (MLC) [2], [5], [22]. In the studies of [2] and [5], the extraction of visual and semantic features shares the same network, but in practice it is found that the MLC task is easier to learn than text generation; the shared network tends to learn the former task with higher priority and incurs lower loss, whereas the latter task performs worse. Hence, the proposed encoder selects two separate CNN branches, i.e., the CNN backbone and the MLC branch, which are used to extract visual and semantic features, respectively.

1) Visual Feature Extraction

The CNN branch (shown in Figure 2) is responsible for visual feature encoding. For one patient, there may exist one or more radiology images, representing different views. Yuan

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056175, IEEE Access

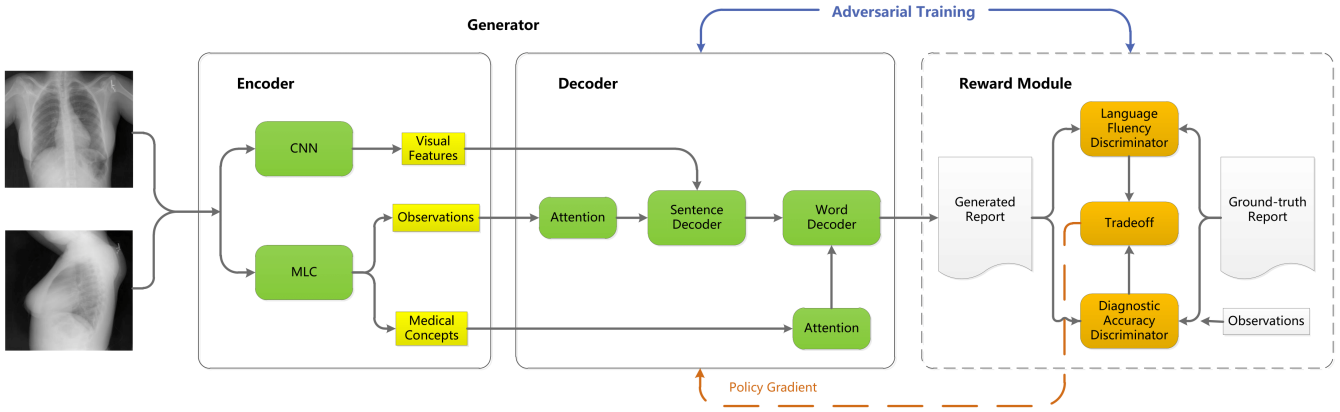


FIGURE 2. **Overview of proposed framework involving three components: encoder, decoder, and reward module. The encoder is comprised of two separate branches that extract visual and semantic features for the decoder separately. The decoder generates a report hierarchically. Note that the decoding process is not unfolded due to space limitations. In adversarial training, the generator is trained by RL using rewards offered by the discriminators. Simultaneously, the discriminators are updated by the maximum-likelihood estimation approach.**

et al. [5] conducted a multi-view fusion on deep features after the CNN backbone. However, an implicit and simpler approach is adopted here; that is, different views are treated as different input channels and it is assumed that the deep CNN can learn how to perform effective feature fusion during training.

In the proposed approach, the CNN backbone is roughly based on Resnet-152 [23] for the following reasons. First, it has a deeper structure, which is suitable for feature representation learning. Second, the CNN module is at the bottom of the generative model, and it is easy to suffer the vanishing gradient problem. However, the residual block of ResNet is helpful to solve this problem. The last classification layer is removed and an adapter module added for channel number conversion at both ends. This adapter module is composed of a convolutional layer, batch-normalization layer, and rectified linear unit (ReLU).

Given a group of frontal and lateral radiology images, they are resized and packed into a three-dimensional (3D) tensor $\mathbf{I}$ with the shape of $224 \times 224 \times 2$, which means that each channel represents a specific view. Then, the input tensor is fed into the proposed CNN backbone and finally encoded to a 1024-dimensional visual vector $\mathbf{v}$.

2) Semantic Feature Extraction

The MLC branch maps raw images to intermediate semantic labels that can help the rear decoder generate words more accurately. In the proposed approach, a submodel is used to predict medical label probabilities for given images; this probability vector and label embedding are viewed as semantic features that will serve as the input of the rear decoder.

Following [5], the semantic information is derived from two sources: 14 common chest radiographic observations [24] (including *enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, and support devices* and *no finding*) and MeSH labels. MeSH is a standard medical vocabulary used for indexing, cataloging, and searching of biomedical and health-related information. These two types of labels provide information from different semantic granularities. The observation labels focus on generalized medical concepts, while the MeSH terms narrow these concepts to medical keywords. The total number of MeSH labels depends on the dataset, so it is denoted M.

VGG-16 [25] is a common CNN backbone with strong fitting ability. An improvement of VGG-16 is to use several consecutive 3x3 convolution kernels to replace larger convolution kernels (11x11 or 5x5). For a given receptive field, using a stacked small convolution kernel is better than using a large convolution kernel, because multiple non-linear layers can learn more complex patterns and the cost is relatively small. Therefore, it has been widely used in medical image classification [26]–[28].

A small multi-task network, which is based on the VGG-16, is built to predict these two types of labels by MLC. The two tasks share the same backbone with different classification layers. For observation prediction, given the image tensor $\mathbf{I}$, the MLC network outputs a 14-dimensional vector $\mathbf{l}^o = (l_1^o, l_2^o, \ldots, l_{14}^o)^{\mathrm{T}}$, where the ith value represents the probability for the ith observation label. The observation MLC task is trained by binary cross-entropy (BCE) loss [26], [29]:

$$L_{MLC}^o = -\sum_{i=1}^{14} \left(w_i^+ l_i^{o\star} \log l_i^o + w_i^- \left(1 - l_i^{o\star}\right) \log \left(1 - l_i^o\right) \right), \tag{1}$$

where $\mathbf{l}^{o\star} = (l_1^{o\star}, l_2^{o\star}, \ldots, l_{14}^{o\star})^{\mathrm{T}}$ denotes the ground-truth binary labels of observations, and w_i^+ and w_i^- are loss weights for the ith class, determined by the class distribution, to handle the data imbalance problem,

$$w_i^+ = \frac{|l_i^{o\star} = 1|}{K},\ w_i^- = \frac{|l_i^{o\star} = 0|}{K}, \tag{2}$$

where K denotes the number of samples and $|\cdot|$ represents the count operation.

Similarly, one can obtain the MeSH loss function L_{MLC}^m. Finally, the total MLC loss L_{MLC} is the weighted sum of the two,

$$L_{MLC} = L_{MLC}^o + L_{MLC}^m. \tag{3}$$

C. DECODER

The proposed decoder is composed of hierarchical LSTMs and multi-level attention. Given encoded features, the sentence LSTM generates topic vectors in an external loop, and the word LSTM outputs words in an internal loop starting with one topic vector (shown in Figure 3). Moreover, it is worth noting that the attention mechanism is used in this procedure. The extracted semantic information, including observations and MeSH predictions, is used with the hidden state of the sentence and word LSTM, respectively, helping to generate abnormal-aware reports.

1) Sentence Decoder

As shown in Figure 3, the sentence LSTM is initialized with the encoded visual vector $\mathbf{v}$, and the input is derived from the attention between the hidden state and observation embedding weighted by their corresponding probabilities. The topic vector and stop probability are mapped from the hidden state.

Similar to [5], the attention function $\mathbf{s}_{att}^o = \mathrm{ATTEN}\left(\mathbf{S}^o, \mathbf{l}^o, \mathbf{h}_s\right)$ is defined as follows:

$$\alpha = \mathrm{softmax}\left(\mathbf{W}_{att}\tanh\left(\mathbf{S}^o\mathbf{W}_{s,1}\mathbf{l}^o + \mathbf{W}_{s,2}\mathbf{h}_s\right)\right), \mathbf{s}_{att}^o = \mathbf{S}^{o\mathrm{T}}\alpha, \tag{4}$$

where $\mathbf{S}^o \in \mathbb{R}^{14\times D}$ is the observation embedding matrix, $\mathbf{l}^o \in \mathbb{R}^{14}$ are predicted observation probabilities, $\mathbf{h}_s \in \mathbb{R}^D$ is the hidden state of the sentence LSTM, $\alpha \in \mathbb{R}^{14}$ are attention weights, $\mathbf{s}_{att}^o \in \mathbb{R}^D$ is the attention output, and $\mathbf{W}_{att} \in \mathbb{R}^{14\times 14}$, $\mathbf{W}_{s,1} \in \mathbb{R}^{D\times 14}$, and $\mathbf{W}_{s,2} \in \mathbb{R}^{14\times D}$ are learnable parameter matrices. Note that the bias term is omitted in this paper for simplicity.

Given the visual vector $\mathbf{v}$, the observation embedding matrix $\mathbf{S}^o$, and predicted observation probability vector $\mathbf{l}^o$, the decoding recurrence of sentence decoder is listed in Algorithm 1.

In Algorithm 1, $\tau^{(n)} \in \mathbb{R}^{2D}$ denotes the topic vector for the nth sentence, $y_s^{(n)}$ is a real number, representing the probability of stopping recurrence, and $\mathbf{W}_\tau \in \mathbb{R}^{D\times 2D}$ and $\mathbf{W}_s \in \mathbb{R}^{D\times 1}$ are the learnable parameter matrices. $\mathrm{LSTMCell}(\cdot)$ represents the standard operation of LSTM for one step. γ denotes the threshold of stopping generation. For simplicity, the cell state of LSTM is omitted in this paper. The output topic vectors will be used as the initial hidden state of LSTM in the word decoder.

Algorithm 1 Decoding Recurrence of Sentence Decoder

Input: $\mathbf{v}$: visual vector; $\mathbf{S}^o$: observation embedding matrix; $\mathbf{l}^o$: predicted observation probability vector; N_{max}: maximum number of topics. γ: stopping threshold.

Output: topic vectors $\left(\tau^{(1)}, \tau^{(2)}, \dots\right)$

initial $n = 1$, $\mathbf{h}_s^{(0)} = \mathbf{v}$;
repeat
  compute attention $\mathbf{S}_{att}^{o(n)} = \mathrm{ATTEN}\left(\mathbf{S}^o, \mathbf{l}^o, \mathbf{h}_s^{(n-1)}\right)$;
  run one-step LSTM operation $\mathbf{h}_s^{(n)} = \mathrm{LSTMCell}\left(\mathbf{S}_{att}^{o(n)}, \mathbf{h}_s^{(n-1)}\right)$;
  compute a topic vector $\tau^{(n)} = \tanh\left(\mathbf{W}_\tau \mathbf{h}_s^{(n)}\right)$;
  compute stop probability $y_s^{(n)} = \mathrm{sigmoid}\left(\mathbf{W}_s \mathbf{h}_s^{(n)}\right)$;
  $n = n + 1$;
until $y_s^{(n)} > \gamma$ **or** $n > N_{max}$

In the pre-training stage, the decoder is trained using cross-entropy loss. The partial loss from the sentence decoder is

$$L_{sent} = -\sum_{n=1}^{N}\left(y_s^{\star(n)}\log y_s^{(n)} + \left(1 - y_s^{\star(n)}\right)\log\left(1 - y_s^{(n)}\right)\right), \tag{5}$$

where $\mathbf{y}_s^\star = \left(y_s^{\star(1)}, y_s^{\star(2)}, \dots, y_s^{\star(N)}\right)^{\mathrm{T}}$ denotes the ground-truth stopping labels for one report. For example, if one report has four sentences, then $\mathbf{y}_s^\star = (0, 0, 0, 1)^T$.

2) Word Decoder

Similar to the sentence decoder, the word decoder is also based on a single-layer LSTM. The word LSTM initialized with a topic vector is fed with attended MeSH embeddings and previously generated word embeddings. The attention operation is the same as that of the sentence decoder, except that the inputs are replaced by MeSH embeddings $\mathbf{S}^m \in \mathbb{R}^{M\times D}$, MeSH probabilities $\mathbf{l}^m \in \mathbb{R}^M$, and the hidden state of the word LSTM, $\mathbf{h}_w$. The word-decoding recurrence for the nth sentence is demonstrated in Algorithm 2.

In Algorithm 2, the superscript (n, t) denotes that the current sentence step is n and the word step is t; $< bos >$ and $< eos >$ denote the beginning and end tokens of one sentence, respectively. $\mathbf{W}_e \in \mathbb{R}^{D\times D}$, $\mathbf{W}_m \in \mathbb{R}^{D\times D}$, and $\mathbf{W}_h \in \mathbb{R}^{D\times C}$ are learnable parameter matrices. $\mathrm{Embedding}$ represents the embedding layer and $\mathbf{y}_w^{(n,t)} \in \mathbb{R}^C$ the predicted word distribution. $\mathrm{sampling}$ denotes the operation of sampling a word from the probability distribution $\mathbf{y}_w^{(n,t)}$. Finally, a full report is obtained by collecting all sampled words in order.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056175, IEEE Access

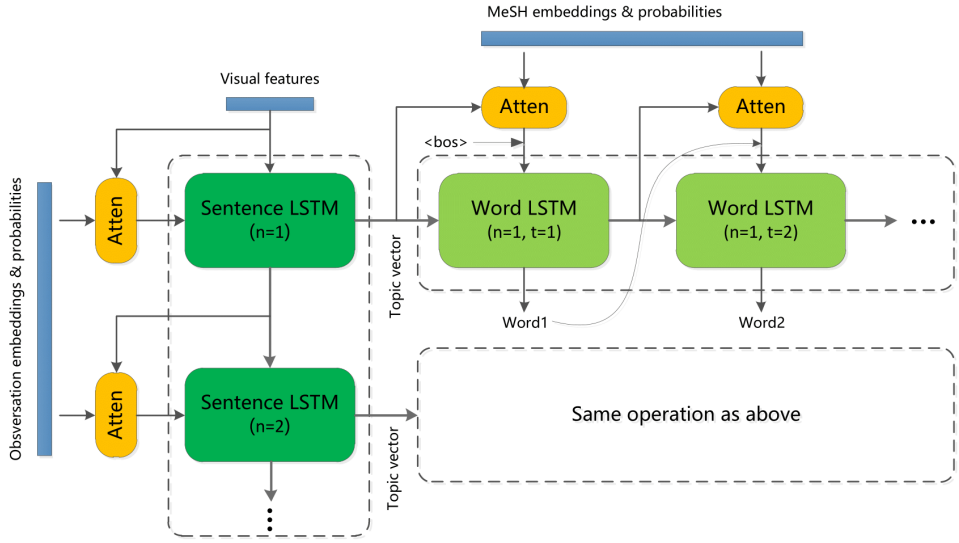


FIGURE 3. **Workflow of hierarchical decoder. For simplicity, the embedding layer in word LSTM is not shown. In general, there exist two-level loops: the sentence LSTM unfolds in the vertical direction and the word LSTM in the horizontal direction.**

Algorithm 2 Decoding Recurrence of Word Decoder

Input: $\tau^{(n)}$: nth topic vector; $\mathbf{S}^m$: MeSH embedding matrix; $\mathbf{l}^m$: predicted MeSH probability vector; T_{max}: maximum number of words per sentence.

Output: nth sentence $\left(z^{(n,1)}, z^{(n,2)}, \ldots\right)$

initial $t = 1$, $z^{(n,0)} =< bos >$, $\mathbf{h}_w^{(n,0)} = \tau^{(n)}$;
repeat
  compute attention $\mathbf{S}_{att}^{m(n,t)} = \mathrm{ATTEN}\left(\mathbf{S}^m, \mathbf{l}^m, \mathbf{h}_w^{(n,t-1)}\right)$;
  compute input $\mathbf{a}^{(n,t)} = \mathbf{W}_e \mathrm{Embedding}\left(z^{(n,t)}\right) + \mathbf{W}_m \mathbf{S}_{att}^{m(n,t)}$
  run one-step LSTM operation $\mathbf{h}_w^{(n,t)} = \mathrm{LSTMCell}\left(\mathbf{a}^{(n,t)}, \mathbf{h}_w^{(n,t-1)}\right)$;
  compute word probability $\mathbf{y}_w^{(n,t)} = \mathrm{softmax}\left(\mathbf{W}_h \mathbf{h}_w^{(n,t)}\right)$;
  sample next word $z^{(n,t)} = \mathrm{sampling}\left(\mathbf{y}_w^{(n,t)}\right)$;
  $t = t + 1$;
until $z^{(n,t)} =< eos >$ **or** $t > T_{max}$

In the pre-training state, the loss of the word decoder is

$$L_{word} = -\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{C} y_{w,i}^{\star(n,t)} \log y_{w,i}^{(n,t)} = -\sum_{n=1}^{N}\sum_{t=1}^{T} \log y_{w,gt}^{(n,t)}, \tag{6}$$

where $\mathbf{y}_w^{\star(n,t)} = \left(y_{w,1}^{\star(n,t)}, y_{w,2}^{\star(n,t)}, \ldots, y_{w,C}^{\star(n,t)}\right)^{\mathrm{T}}$ denotes the one-hot encoded ground truth, C is the size of vocabulary, and $y_{w,gt}^{(n,t)}$ represents the probability retrieved with an index of the ground-truth token in $\mathbf{y}_w^{(n,t)}$ (i.e., the likelihood for the ground-truth token).

The total loss of the decoder in the pre-training stage is the weighted sum of L_{sent} and L_{word}:

$$L_g = w_{sent} L_{sent} + w_{word} L_{word}. \tag{7}$$

D. REWARD MODULE

The reward module composed of two discriminators provides rewards for the generator in ARL. In this section, details of the reward module design are given.

To fully measure the quality of a generated medical image, two aspects are considered: language fluency and diagnostic accuracy. Hence, two discriminators are designed to separately evaluate one report from these two aspects.

1) Readability

Previous work [3], [13], [30], [31] used natural language generation (NLG) metrics like BLEU [32], ROUGE [33], and CIDEr [34] as important metrics with which to calculate the reward. However, this rule-based method cannot capture natural language structure information and can often be misleading. For example, repeated words or sentences can simply improve these metrics, although it has no practical use.

A language fluency discriminator D_f is designed to measure the fluency of a generated report. Given an embedded word sequence of a medical report $\mathbf{X} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\right)$, the following formulation is used to calculate the fluency reward:

$$r_f = D_f\left(\mathbf{X}\right) = \mathrm{sigmoid}\left(\mathbf{W}_f \mathrm{LSTM}\left(\mathbf{X}\right)\right), \tag{8}$$

where r_f is the fluency reward for the given report. A higher value means better language fluency. $\mathbf{W}_f \in \mathbb{R}^{D\times 1}$ is a learnable parameter matrix, and LSTM in the above equation

denotes multi-step LSTM operation.

2) Accuracy

To measure the diagnostic accuracy, an accuracy discriminator D_a is designed that checks how much a report can cover the key observations. Note that these observations are parts of labels used in semantic information extraction. D_a takes in embedded word sequence of an entire report $\mathbf{X}$ and a binary vector of observation ground truth $\mathbf{l}^{o\star}$, and outputs a consistency reward r_a. The detailed formulation of D_a is

$$\begin{aligned} r_a &= D_a\left(\mathbf{X}, \mathbf{l}^{o\star}\right) \\ &= \operatorname{sigmoid}\left(\mathbf{W}_{a,o}\left(\left(\mathbf{W}_{a,l}\mathbf{l}^{o\star}\right) \odot \left(\mathbf{W}_{a,h}\mathrm{LSTM}\left(\mathbf{X}\right)\right)\right)\right), \end{aligned} \quad (9)$$

where $\mathbf{W}_{a,h} \in \mathbb{R}^{D\times 14}$, $\mathbf{W}_{a,l} \in \mathbb{R}^{14\times 14}$, $\mathbf{W}_{a,o} \in \mathbb{R}^{14\times 1}$ are parameter matrices of the accuracy discriminator.

3) Reward Function

Considering both language fluency and diagnostic accuracy, the weighted sum of r_f and r_a is used as the final reward:

$$\begin{aligned} r\left(\mathbf{X}, \mathbf{l}^{o\star}\right) &= \lambda r_f + (1-\lambda)\, r_a \\ &= \lambda D_f\left(\mathbf{X}\right) + (1-\lambda) D_a\left(\mathbf{X}, \mathbf{l}^{o\star}\right), \end{aligned} \quad (10)$$

where λ is a trade-off parameter within $[0, 1]$.

E. ADVERSARIAL REINFORCEMENT LEARNING

Before ARL, some pre-training steps must be taken: (1) pre-train the embedding layer by word2vec [35], [36]; (2) train the MLC model using loss function 3; (3) fix the MLC and pre-train the encoder-decoder with loss function 7; and (4) pre-train the discriminators with the ground-truth reports and generated reports. The ARL stage is focused on first.

In each training iteration of ARL, the two discriminators are first updated by the loss functions defined later with the fixed generator, and then the generator is optimized using RL by fixing the discriminators (see Figure 4).

1) Training Discriminators

Given generated reports and real reports with their observation labels, discriminators are trained by maximum likelihood estimation, aiming to give a higher reward for real data and lower for generated data.

The one positive sample pair is defined as $(\mathbf{I}, \mathbf{X}^{\star}, \mathbf{l}^{o\star})$, where $\mathbf{I}$ denotes image data, and $\mathbf{X}^{\star}$ and $\mathbf{l}^{o\star}$ represent its corresponding ground-truth report sequence and observation labels, respectively. Feeding image data into the generator, one can obtain a negative sample pair $(\mathbf{I}, \mathbf{X}, \mathbf{l}^{o\star})$. Note that the ground-truth observation labels are still used instead of the predictions by MLC. According to maximum-likelihood estimation, the losses of FD and AD comprise the negative log-likelihood:

$$L_{fd} = -\log D_f\left(\mathbf{X}^{\star}\right) - \log\left(1 - D_f\left(\mathbf{X}\right)\right), \quad (11)$$

$$L_{ad} = -\log D_a\left(\mathbf{X}^{\star}, \mathbf{l}^{o\star}\right) - \log\left(1 - D_a\left(\mathbf{X}, \mathbf{l}^{o\star}\right)\right). \quad (12)$$

2) Training Generator

Given the reward provided by discriminators, the generator is refined by RL to generate more realistic reports. The generator is viewed as an agent, and the input image and previously generated words as environmental states s. An "action" a refers to the prediction of the next word conditioned on the input image and previous actions in decoding recurrence. The agent interacts with the environment and takes actions based on its policy π_θ, which is defined by the parameters in the generator.

In the training process, the policy π_θ is performed stochastically to sample a word at every time step and uses it as input for the next time step. Stochastic policy represents that the probability distribution of the agent choosing an action equals the predictive probability distribution by $\pi_\theta(a \mid s)$:

$$\pi(a \mid s) \triangleq p(a \mid s), \quad (13)$$

$$\sum_{a\in\mathcal{A}} \pi(a \mid s) = 1, \quad (14)$$

where a denotes an action belonging to an action space $\mathcal{A}$, and s is a description of the environment.

The final reward r can be calculated by the reward module after all words are generated. The training goal is to minimize the negative expected reward:

$$L_g^{rl}(\theta) = -\mathbb{E}_{\mathbf{X}\sim\pi_\theta}\left[r\left(\mathbf{X}, \mathbf{l}^{o\star}\right)\right], \quad (15)$$

where $\mathbf{X}$ denotes the sampled report, $\mathbf{l}^{o\star}$ ground-truth observation labels, and $r\left(\mathbf{X}, \mathbf{l}^{o\star}\right)$ the total reward provided by discriminators for this sampled report. According to the REINFORCE algorithm [37], the gradient of $L_g^{rl}(\theta)$ with respect to θ can be represented as

$$\begin{aligned} \nabla_\theta L_g^{rl}(\theta) &= -\mathbb{E}_{\mathbf{X}\sim\pi_\theta}\left[r\left(\mathbf{X}, \mathbf{l}^{o\star}\right) \cdot \nabla_\theta \log \pi_\theta\left(\mathbf{X}\right)\right] \\ &\approx -r\left(\mathbf{X}, \mathbf{l}^{o\star}\right) \cdot \nabla_\theta \log \pi_\theta\left(\mathbf{X}\right). \end{aligned} \quad (16)$$

To reduce the variance of gradients, a baseline is subtracted from the origin reward, and the baseline reward is derived from the greedy decoding of the generator similarly to the self-critical algorithm [14]:

$$\nabla_\theta L_g^{rl}(\theta) \approx -\left(r\left(\mathbf{X}, \mathbf{l}^{o\star}\right) - r\left(\hat{\mathbf{X}}, \mathbf{l}^{o\star}\right)\right) \cdot \nabla_\theta \log \pi_\theta\left(\mathbf{X}\right), , \quad (17)$$

where $\hat{\mathbf{X}}$ represents the greedily decoded report, while $\mathbf{X}$ is the sampled report. With the gradient formula given above, one can use the policy gradient (PG) [37] to update the parameter θ. The PG is a gradient-based optimization method for RL, and it can be used even if the agent is a continuously differentiable function. In the training process, if a series of actions achieve a high reward, the PG will update the parameters to improve the joint probability of these actions.

In conclusion, the ARL algorithm is detailed in Algorithm 3.

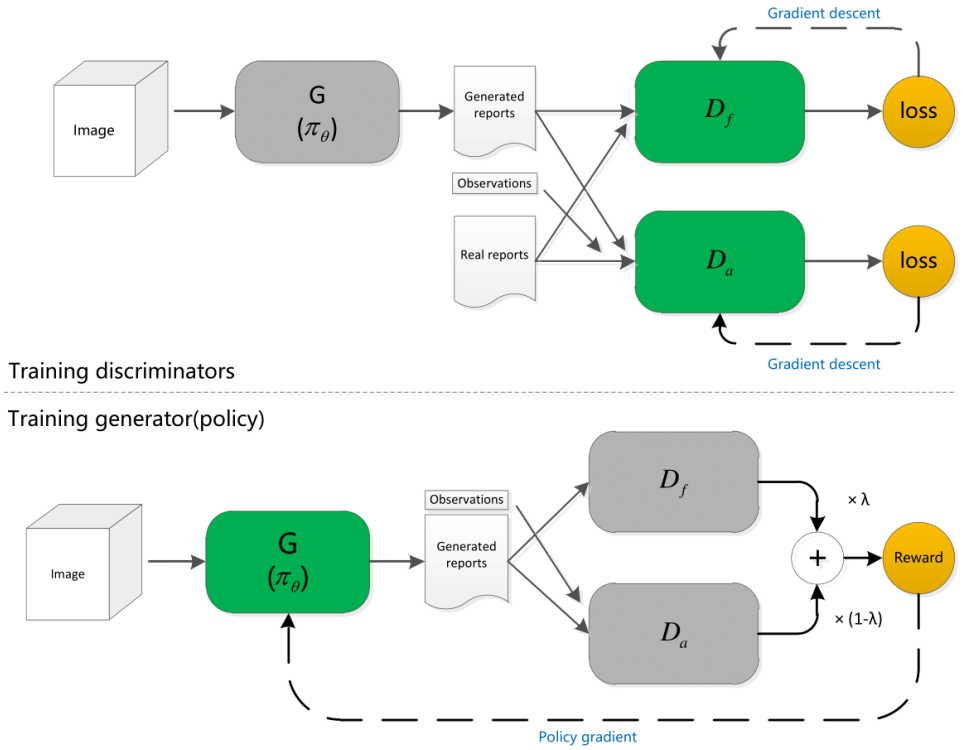


FIGURE 4. **Illustration of adversarial training. The upper part of the figure shows the process of training the generator, and the lower part demonstrates that of training the discriminators. In this figure, "G" refers to the generator (also the policy), D_f to the fluency discriminator, and D_a to the accuracy discriminator. The parameters of gray blocks are not learnable. D_a takes in a report (real or generated) and its corresponding observation ground truth to offer an accuracy reward by comparing the consistency of each other, while D_f "reads" a report to provide a fluency reward that is used to measure how likely it has been written by a human expert.**

Algorithm 3 ARL algorithm

Input: $\mathbf{I}$: image tensor; $\mathbf{X}^\star$: embedded ground-truth report sequence; $\mathbf{l}^{o\star}$: ground-truth observation vector; λ: trade-off parameter.

Output: the optimal D_f, D_a, π_θ

repeat
  sample one data pair $(\mathbf{I}, \mathbf{X}^\star, \mathbf{l}^{o\star})$ from training set;
  sample report based on policy $\mathbf{X} \sim \pi_\theta(\cdot \mid \mathbf{I})$;
  compute fluency rewards for real and generated reports $D_f(\mathbf{X}^\star)$, $D_f(\mathbf{X})$;
  compute accuracy rewards for real and generated reports $D_a(\mathbf{X}^\star, \mathbf{l}^{o\star})$, $D_a(\mathbf{X}, \mathbf{l}^{o\star})$;
  optimize D_f and D_a by loss function (12);
  compute $D_f(\mathbf{X})$, $D_a(\mathbf{X}, \mathbf{l}^\star)$ again with updated models, and compute total reward $r(\mathbf{X}, \mathbf{l}^{o\star})$ by equation (10);
  greedily sample $\hat{\mathbf{X}} \sim \pi_\theta(\cdot \mid \mathbf{I})$, and compute its total reward $r\left(\hat{\mathbf{X}}, \mathbf{l}^{o\star}\right)$
  optimize policy π_θ by gradient function (17);
until D_f,D_a,π_θ converge

IV. EXPERIMENTS

In this section, the datasets used, implementation details, compared methods, and evaluation metrics are presented. Finally, experimental results and discussion are provided.

A. DATASETS

IU X-Ray. The Indiana University Chest X-ray Collection [38] is a public dataset consisting of 7,470 frontal and lateral view images paired with their corresponding 3,955 diagnostic reports. Each report contains four parts: comparison, indication, findings, and impression. Our focus is generating the findings in this paper. To acquire the ground truth for observations and MeSH, CheXpert labeler [1] [24] and NIH MTI web API [2], respectively, are utilized. Given a report, the first tool gives four possible values for each of the 14 observations: present (1), absent (0), uncertain (-1), or not mentioned (blank), and then it is mapped to a binary label (replacing -1 with 1 and filling 0 in blanks). The second tool extracts several keywords as MeSH labels from origin reports. The minimal occurrences are empirically set as 20 and 50 MeSH labels obtained. Then, the image-report pairs are selected by the following conditions: (1) the report of the sample corresponds to multiple views; (2) the report contains the finding part; (3) the report has more than three sentences; and (4) the sample is assigned to at least one observation and MeSH label. After selection, 2,658 reports associated with

[1]https://github.com/stanfordmlgroup/chexpert-labeler
[2]https://ii.nlm.nih.gov/MTI/index.shtml

TABLE 1. Observations and MeSH term occurrences on datasets; only the top 14 MeSH terms are listed.

Dataset	Observation	Count	MeSH	Count
IU X-Ray	No Finding	1243	Lung	1638
	Lung Opacity	602	Mediastinum	1036
	Cardiomegaly	514	Heart	556
	Enlarged Cardiomediastinum	329	Thorax	409
	Atelectasis	243	Lung Diseases	362
	Support Devices	177	Pulmonary Atelectasis	240
	Lung Lesion	149	Cardiomegaly	157
	Pleural Effusion	131	Pulmonary Edema	151
	Fracture	109	Aorta	128
	Edema	51	Edema	121
	Consolidation	50	Granuloma	117
	Pneumonia	48	Pneumonia	115
	Pneumothorax	47	Bone and Bones	110
	Pleural Other	42	Diaphragm	93
MIMIC-CXR	No Finding	9736	Lung	11275
	Lung Opacity	3440	Mediastinum	6985
	Atelectasis	3011	Exudates and Transudates	6362
	Pleural Effusion	2969	Thorax	5795
	Pneumonia	2897	Pulmonary Atelectasis	4120
	Cardiomegaly	2320	Pulmonary Edema	3591
	Edema	2027	Lung Diseases	2829
	Support Devices	1571	Pneumonia	2632
	Consolidation	675	Edema	1886
	Lung Lesion	647	Diaphragm	1615
	Pneumothorax	470	Cardiomegaly	1567
	Enlarged Cardiomediastinum	468	Heart	1477
	Fracture	462	Pleura	1259
	Pleural Other	290	Aorta	1253

5,316 images are obtained. In the tokenization process, the low-frequency tokens with occurrences lower than three are ignored and, finally, 889 tokens in total are yielded.

MIMIC-CXR. MIMIC-CXR is a large publicly available dataset of labeled chest radiographs [39] that provides more than 300,000 chest x-rays associated with corresponding medical reports. Data selection is the same as the first dataset. In the experiments reported here, only a part of the dataset is selected because of limited computing resources; nevertheless, MIMIMC-CXR is 7 times larger than the first dataset, reaching 19,364 pairs. The observation labels are provided by this dataset and the MeSH labels with a number of 120 are achieved in the same way as above. Finally, the report texts are tokenized to 2,104 unique words with a minimum frequency of five.

The rank of occurrences for observation and MeSH labels on the two datasets is listed in Table 1. From the table, it can be found that the MeSH terms are more specific on medical concepts in contrast with the observations. In experiments, both datasets are split into training, validation, and test sets according to the ratio 7/1/2.

Three measures are taken to tackle the data imbalance problem. First, the normal cases are under-sampled and their proportion reduced to less than 60%. Second, a stratified sampling approach is adopted to split the training/validation/test set to keep the proportion roughly equal in different sets. Finally, the weighted BCE loss is used in the optimization of the MLC model.

B. IMPLEMENTATION DETAILS

All proposed models are implemented on PyTorch [40]. First, the embedding layer is pre-trained by gensim [36] with the total reports of the two datasets separately; gensim is a python module that implements the word2vec [35] family of algorithms, using highly optimized C routines. The embedding dimension D is set to 512.

TABLE 2. Complexity of proposed model.

	Model	Parameters	FLOPs
Encoder	CNN(ResNet-152)	60,244,041	11,559,341,056
	MLC(VGG-16)	221,075,730	15,594,068,992
Decoder	SentDecoder	2,642,160	8,215,116
	WordDecoder	4,011,419	1,309,203,420
Reward	AD	2,108,655	631,610,578
	FD	2,101,761	631,603,712
Total		292,183,766	29,734,042,874

The MLC backbone is based on VGG-16 [25], and the origin classifier is replaced by the two classifiers proposed for observations and MeSH terms. In MLC training (the second training state), the maximum epoch is set to 200, Adam [41] is adopted as the optimizer, with a learning rate of 0.0005, and the batch size set to 20. Early stopping is taken when the validation loss does not decrease for 5 epochs. The dimension of LSTM in the encoder-decoder equals 512 and the dimension of topic vector 1,024.

In the third pre-training stage, the CNN backbone is roughly based on ResNet-152 [23]. The parameters of the MLC are fixed and the encoder-decoder model pre-trained by cross-entropy loss. The Adam optimizer is also adopted, but the learning rate is set to 0.00005. The training process is stopped if the CIDEr of the validation set does not rise for 5 epochs.

The generator is used to sample "fake" reports, and these reports, with ground-truth reports and observations, are viewed as the dataset with which to pre-train the discriminators. The optimization method used is also Adam with a learning rate of 0.0005. The training process is stopped if the validation loss does not decrease.

In the ARL stage, the tradeoff parameter is set to 0.5 and the batch size changed to 5. The SGD optimizer is used with a momentum of 0.9, and the learning rate is 0.00005. Early stopping is also used.

In inference mode, the maximum number of sentences per report N_{max} is 10 and the maximum number of words per sentence T_{max} is 50. The stopping threshold γ is 0.5.

Table 2 illustrates the complexity of each module in the proposed model. Two common measures are adopted: the number of learning parameters and the floating-point operations (FLOPs). Note that some values in this table are related to the size of vocabulary or the length of a report, and these values are calculated based on the worst case.

C. COMPARED METHODS

To verify the performance of the proposed approach, four competitive models are chosen as the compared methods.

CNN-RNN [10] is a canonical image-captioning model proposed in 2015, composed of a CNN as the encoder and a single-level RNN as the decoder. For comparability, the encoder is the same as the proposed CNN model and the decoder is the same as the proposed word decoder. The model settings are also consistent. Additionally, this model is trained by minimizing cross-entropy loss.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056175, IEEE Access

CoAtt [2] Different from the CNN-RNN, the CoAtt model introduced the co-attention and hierarchical RNN to the decoder. In the encoder, the visual features and semantic features (embedded MeSH terms) are extracted from the same network, and these features are only combined with sentence LSTM for the topic generation. In experiments, the model is trained by minimizing the cross entropy loss produced in stopping and word prediction.

MvH-AttL-MC [5] The MvH-AttL-MC inference model is similar to the proposed model. The encoder encodes visual features and medical concept features, and the decoder is hierarchical with multi-level attentions. However, the differences of inference architecture are (1) visual features and medical information share the same backbone similar to CoAtt, and (2) the predicted observations are not fused into the attention mechanism in the decoder. In addition, this model is also trained by minimizing the cross-entropy.

NLGR-CCR [3] The NLGR-CCR model adopts a deep CNN as the encoder and a hierarchical LSTM as the decoder. More importantly, it is fine-tuned by RL with two empirically designed rewards, i.e., NLG and CCR.

Since the authors of CoAtt and MvH-AttL-MC did not release their source code, their models were implemented based on details in their papers. Some implementation details and hyper-parameters may be different from the original settings, and, accordingly, some gaps exist between the experimental results.

D. EVALUATION METRICS

To fully evaluate the performance of the proposed model, several different metrics considering both language fluency and diagnostic accuracy are adopted.

First, popular natural language generation (NLG) evaluation metrics, including BLEU-4 [32], METEOR [42], ROUGE-L [33], and CIDEr [34], are adopted, which are used to measure the statistical correlation between two text sequences. In the experiments, the automatic tool [3] is used to calculate these metrics.

Second, to measure the diagnostic accuracy of generated reports, the CheXpert labeler [24] is applied to the generated reports, and the precision, recall, and F1 scores are calculated.

E. PERFORMANCE COMPARISONS

In this section, first all the models are separately trained on the two datasets. The training sets are used for parameter learning, validation sets for early stopping, and the test sets are invisible during training for conducting experiments. To ensure the comparability of experimental results, all the hyper-parameters are kept as consistent as possible.

1) NLG Metric Evaluation

The performance comparisons by NLG metrics on both datasets are presented in Table 3. In general, the full proposed model outperforms all baselines across all metrics. As expected, the CNN-RNN performs the worst among all models, especially for the MIMIC-CXR dataset, which indicates that a simple encoder-decoder architecture without techniques has limited learning ability when used on large amounts of data. On the IU X-Ray dataset, the NLG metric scores are very close within CoAtt, MvH-AttL-MC, NLGR-CCR, and the proposed model, even though the latter model's scores are better by 1%–2%. The narrow gap may be due to three reasons: (1) the data volume of IU X-Ray is relatively small, which limits the representation capability of these models; (2) the sentence structure of the medical report is relatively simple; and (3) the average length of reports is relatively short. In contrast, the gaps between different methods are obvious on the larger MIMIC-CXR dataset. The proposed model achieves the best performance across all metrics, leading the second-best performer, NLGR-CCR, by 4%-5%, and the two methods are both retuned by RL. Despite the similar inference architecture, the proposed model clearly outperforms on MvH-AttL-MC, which indicates that the improvement in training scheme (employing ARL) can help generate highly relevant finding reports.

TABLE 3. NLG metric scores(%) of different methods on test sets from IU X-Ray and MIMIC-CXR datasets. The complete proposed architecture is labeled "Full-ARL" in contrast with the variants (decomposed models) in the follow-up study.

Dataset	Model	BLEU-4	METEOR	ROUGE-L	CIDEr
IU X-Ray	CNN-RNN	6.03	12.4	20.6	23.2
	CoAtt	9.21	13.7	23.4	30.5
	MvH-AttL-MC	11.4	16.2	24.1	32.3
	NLGR-CCR	10.2	15.3	25.3	34.7
	Full-ARL	**12.5**	**17.1**	**26.2**	**36.6**
MIMIC-CXR	CNN-RNN	4.12	12.1	19.7	21.5
	CoAtt	9.23	14.3	22.6	31.4
	MvH-AttL-MC	10.4	17.8	23.5	33.8
	NLGR-CCR	10.5	22.1	27.3	37.6
	Full-ARL	**14.8**	**25.3**	**32.9**	**40.2**

2) Diagnostic Accuracy Evaluation

For medical report generation, a more important metric is diagnostic accuracy. Similar to [3], the CheXpert [24] labeler is utilized to achieve the binary labels from the generated reports and ground-truth reports for 14 observations, and then the precision, recall, and F1 scores are computed for each observation class. The class distribution of the IU X-Ray dataset is extremely unbalanced, so only an evaluation of the full proposed model is conducted on the MIMIC-CXR dataset and compared with MvH-AttL-MC and NLGR-CCR.

Table 4 illustrates the detailed comparison results of diagnostic accuracy. Generally, the proposed model achieves the highest scores in most observation classes. Specifically, for the high-frequency observations, e.g., *No Finding*, the proposed model can maintain the balance between precision and recall, while the other two methods prefer the recall score. Even though the diagnostic accuracy metric is susceptible to observation proportion, the proposed model obtains the highest F1 score on the small-share observations. In

[3]https://github.com/tylin/coco-caption

TABLE 4. Diagnostic accuracy evaluation results of different methods on MIMIC-CXR dataset. Second column lists the proportion (%) of each observation, and starting from the next column, it exhibits precision/recall/F1 (%) for different methods across all observations, where "-" denotes that precision and recall equal zero, and thus F1 score is undefined. Note that the word "Enlarged" in the observation column is short for "Enlarged cardiomediastinum" since this phrase is too long to display in the table.

Observation	Proportion	MvH-AttL-MC	NLGR-CCR	Full-ARL
No Finding	50.2	60.0/**92.4**/72.7	71.2/87.8/**78.6**	**71.8**/74.1/72.9
Lung Opacity	17.8	39.9/23.2/29.3	**53.0**/38.5/44.6	52.2/**45.8/48.8**
Atelectasis	15.5	34.5/11.6/17.4	39.7/21.4/27.8	**42.6/25.9/32.2**
Pleural Effusion	15.3	26.7/7.61/11.9	**35.2**/12.0/17.9	34.7/**14.0/20.0**
Pneumonia	15	25.4/7.18/11.2	33.7/11.7/17.3	**35.6/14.8/20.9**
Cardiomegaly	12	17.2/4.44/7.05	20.3/6.29/9.61	**22.1/7.15/10.8**
Edema	10.5	8.50/0.84/1.53	12.7/2.81/4.60	**13.4/3.85/5.98**
Support Devices	8.11	6.06/0.51/0.94	11.54/1.15/2.08	**14.3/1.78/3.17**
Consolidation	3.49	-	5.75/0.74/1.31	**8.99/1.19/2.09**
Lung Lesion	3.34	-	4.00/0.31/0.57	**6.67/0.62/1.13**
Pneumothorax	2.43	-	-	**3.23/0.21/0.40**
Enlarged	2.42	-	-	-
Fracture	2.4	-	-	-
Pleural Other	1.5	-	-	-
macro		15.6/10.6/10.9	20.5/13.0/14.6	**21.8/13.5/15.6**
micro		51.9/34.6/41.5	**58.3/36.9/45.2**	56.3/34.5/42.8

addition, among the three methods, MvH-Att-MC performs the worst, which can be reasonably explained by the fact that this model is only optimized by cross-entropy loss and does not consider clinical accuracy. Furthermore, compared to the NLGR-CCR, the proposed model performs better on low-frequency observations, which verifies that its learnable accuracy discriminator can provide effective rewards superior to the rules-based approach. In addition, from the last two rows of Table 4, the proposed model performs best on macro precision, recall, and F1 score, while it is outperformed by NLGR-CCR on micro scores. One possible reason for this is that the micro scores are prone to being influenced by larger-scale classes. Therefore, the macro scores are more reliable and comprehensive in terms of reflecting the performance of all models on diagnostic accuracy.

F. STUDY ON BACKBONES

To explore the effect of the CNN and MLC backbones on the model performance, the CNN and MLC backbones are switched and the output of the MLC and generator evaluated without the discriminators. The experimental results are listed in Table 5, in which the AUC score is used to measure the output of the MLC branch, so it is independent of the CNN backbone. From this column, it can be concluded that VGG-16 is superior to ResNet-152. As shown in Table 2, the number of parameters of VGG-16 is more than twice that of ResNet-152. The other two metrics (CIDEr and F1) are for the generator. As expected, ResNet-152 as the CNN backbone combined with VGG-16 as the MLC branch achieves the best performance.

G. STUDY ON DISCRIMINATORS

In the preceding section, the competitive performance of the proposed full model with different metrics is demonstrated. Next, to verify the effectiveness of the discriminators in the proposed full model, the FD and AD are decomposed separately and ablation experiments conducted on the MIMIC-CXR dataset using metrics similar to those described above. As stated in subsection IV-B, the discriminators must be pre-trained before ARL, and the results on the validation set of the MIMIC-CXR dataset are illustrated in Figure 5.

TABLE 5. Effect of CNN and MLC backbones on model performance.

CNN	MLC	AUC	CIDEr	F1
ResNet-152	VGG-16	**64.6**	**31.5**	**11.2**
VGG-16	VGG-16	**64.6**	29.7	10.4
ResNet-152	ResNet-152	62.1	30.4	9.8
VGG-16	ResNet-152	62.1	28.8	8.7

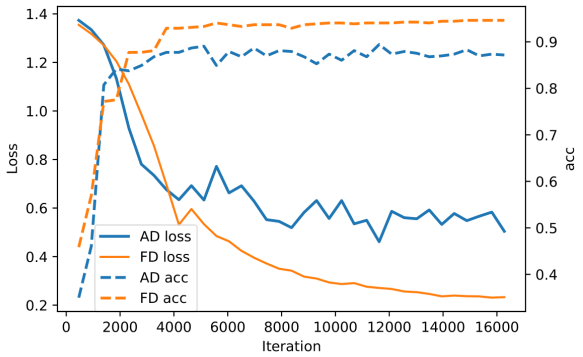


FIGURE 5. **Pre-training for discriminators.**

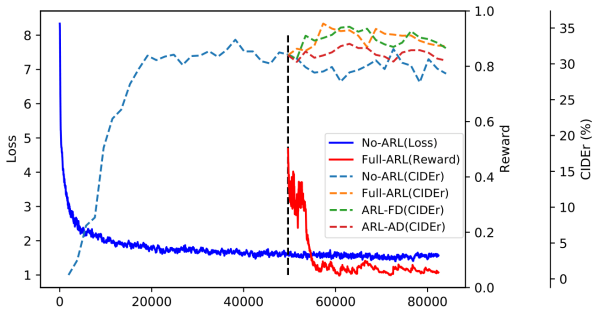


FIGURE 6. **Curves of CIDEr score (%) for variants of proposed model during training. No-ARL denotes that ARL is not used throughout the training process. Full-ARL means that the proposed model is retuned by the full reward module including FD and AD based on No-ARL at the retuning point. ARL-FD and ARL-AD, as their names imply, only adopt one of the two discriminators, respectively, during the retuning state (after black dotted line).**

The comparison of CIDEr curves of different variant models during training is given in Figure 6. The reward shown in this figure is the relative total reward, i.e., the original reward minus the baseline reward. The original rewards are given for sampled reports, while the baseline rewards are given for greedily decoded reports. At the beginning of training, there is a high probability of sampling a report with a higher reward than the greedily decoded report. Therefore, the relative

reward starts with a large value. As the training continues, the policy is trained well, and the original and baseline rewards become very close. Hence, the relative reward is finally close to zero. The loss in the figure is the cross-entropy loss for the generator in the pre-training stage (i.e., the loss of the No-ARL model). The proposed model is first trained by minimizing the cross-entropy loss and computing the CIDEr score on the validation set every 1 epoch. The others (Full-ARL, ARL-FD, and ARL-AD) are initialized by the No-ARL model at the retuning point and trained separately. It can be found from this figure that, without ARL, the CIDEr of the proposed model reaches the upper limit of approximately 33% after 60,000 iterations, and the proposed Full-ARL increases significantly after the turning point. Only with the FD is the CIDEr score 2% higher than that of the No-ARL, while the ARL-AD score does not increase significantly. This fact suggests that the combination of FD and AD can achieve the best performance on textual relevance.

To examine the effectiveness of the discriminators on diagnostic accuracy, a similar experiment with accuracy evaluation is conducted as a performance comparison on variants of the proposed model. The macro precision/recall/F1 scores for observations are shown in Figure 7. The proposed Full-ARL achieves the highest macro recall and F1 scores among all models. The ARL-AD yields the highest macro precision score, while its recall score is slightly lower than that of the proposed full model, which indicates that a single AD may not provide a comprehensive reward and must be assisted with FD to achieve the best performance.

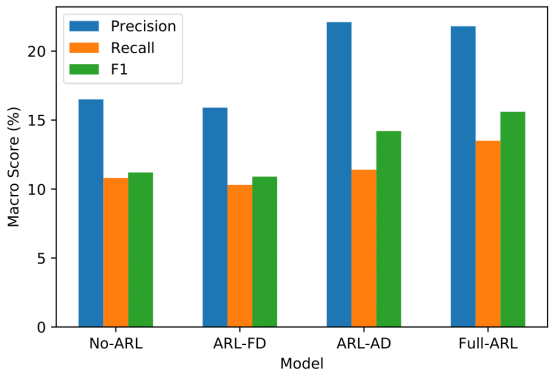


FIGURE 7. Macro precision/recall/F1 scores (%) of different variants of proposed model for observations. Full-ARL results are consistent with those in last row of Table 4.

H. STUDY ON TRADE-OFF PARAMETER λ

In previous experiments, the trade-off parameter λ of Full-ARL is just adopted as 0.5, and in this subsection the sensitivity of the parameter to the performance of the proposed model is explored.

This experiment was only performed on the MIMIC-CXR dataset for the following reasons: (1) larger data size; (2) larger share of abnormal cases; and (3) richer medical vocabulary and longer report paragraphs.

Specifically, λ is varied in the range 0.1–0.9 with a step of 0.2, and the ARL approach conducted based on the No-ARL model separately, to observe the variations of F1 and CIDEr. It is worth noting that there are three cases, i.e., $\lambda = 0, 0.5, 1$, which already been discussed in the study of discriminators.

For better visualization, each dependent variable is rescaled by subtracting the minimum and dividing by the minimum, and the results are illustrated in Figure 8. From the reward equation (10), the parameter λ is the weight of fluency reward and $1 - \lambda$ the weight of accuracy reward. Therefore, one may intuitively assume that the greater the value of λ, the higher the CIDEr (or the smaller the λ and the higher the F1 score). However, as Figure 8 shows, the optimal point of F1 or CIDEr is not at the interval endpoint. For example, the optimal point for F1 is roughly 0.5, while for CIDEr it is approximately 0.7. This result implies that the combination of AD and FD is more likely to "teach" the generator to generate high-quality medical reports.

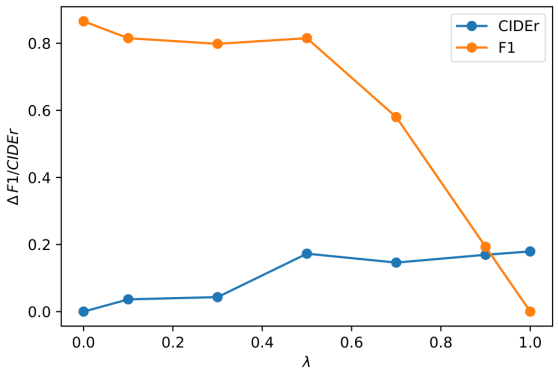


FIGURE 8. Relative change rate of F1 and CIDEr with respect to λ on MIMIC-CXR dataset.

I. QUALITATIVE ANALYSIS

In this subsection, four representative examples are selected for qualitative analysis, as shown in Figure 9. Compared to the proposed full model, No-ARL has limited capability to capture abnormal findings. For example, in the second case, the proposed No-ARL only mentions "heart size is enlarged," while the proposed full model discovers all critical diseases. This proves that the ARL approach can boost abnormal perception and diagnostic accuracy. Moreover, the proposed Full-ARL can generate a longer paragraph than No-ARL, which resembles the work of a radiologist. However, the proposed full model also makes some mistakes. In the third case, it fails to discover "scattered calcified granulomas bilaterally," and thus makes a "No Finding" diagnosis, even though the "granuloma" is among the chosen MeSH labels. In the last case, the proposed full model mentions "atelectasis is seen at the right base," which is not covered in the ground

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056175, IEEE Access

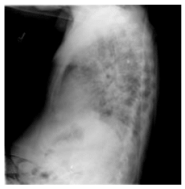
Chest x-ray	Ground truth	Our NoARL	Our FullARL	Label prediction
	heart size within normal limits, stable mediastinal and hilar contours. no focal alveolar consolidation. no definite pleural effusion seen. no typical findings of pulmonary edema.	no active disease. the heart and lungs have in the interval. lungs are clear. no pleural effusion or pneumothorax is seen. no pleural effusion or pneumothorax.	normal heart size and mediastinal contours. there is no alveolar consolidation. there are no findings of pleural effusion. there are no typical findings of pulmonary edema.	No Finding --- Heart, Mediastinal, Pleural, Edema
	**the heart is enlarged**. there is **pulmonary vascular congestion** with diffusely increased interstitial, **mild patchy airspace opacities,** and **pulmonary edema**. there is no pneumothorax or large pleural effusion. there are no acute bony findings.	**heart size is enlarged**. no focal airspace consolidation pulmonary opacity large pleural effusion or pneumothorax. no focal areas of consolidation. lungs are clear.	**the heart size is enlarged.** there is **pulmonary vascular congestion and indistinctness.** there is no evidence of pneumothorax or large pleural effusion. there is **mild pulmonary edema,** interstitial infiltrates difficult to exclude.	Cardiomegaly, Edema, Lung Opacity --- Heart, Enlarged, Pleural, Aorta, Edema
	the heart size and mediastinal contours appear within normal limits. no focal airspace consolidation, pleural effusion or pneumothorax. **scattered calcified granulomas bilaterally.** no acute bony abnormalities.	the heart size is normal. the costomediastinal silhouette is normal. no pneumothorax or pleural effusion. no pleural effusions or pneumothorax.	the lungs are clear without focal consolidation. the mediastinal and hilar contours are normal. the cardiomediastinal silhouette is normal. no acute osseous abnormalities. no pleural effusion or pneumothorax is seen.	No Finding --- Lung, Thorax, Mediastinal, Osseus, Pleural
	**lung volumes are somewhat low.** there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.	the lungs are clear. the lungs are clear. the cardiomediastinal silhouette is normal. no pleural effusion or pneumothorax is seen. no pleural effusion or pneumothorax is seen.	cardiomediastinal contours are normal. the **lung volumes are slightly low**. the size of the **cardiac silhouette is enlarged**, but unchanged from the previous radiograph. the visualized upper abdomen is unremarkable. no pneumothorax, large pleural effusion.	Cardiomegaly --- Lung, Lung Lesion, Abdomen, Pneumothorax, Pleural
	Since ___, a subtle left lower **lobe opacity** has improved and likely represents a **resolving pneumonia**. The scarring in the right middle and left upper lobes is unchanged and fully assessed on CT scan from ___. Two **sclerotic thoracic vertebral bodies,** one with a compression deformity are unchanged.	the lungs are clear without focal consolidation. the lungs are clear. the lungs are clear. the cardiomediastinal silhouette is within normal limits. no pleural effusion or pneumothorax is seen.	as compared to the previous radiograph, the patient has received a right sided picc line. the lungs are clear. there is no pneumothorax. no pleural effusions. no pleural effusions.	Pleural Effusion, Atelectasis --- Lungs, Pneumonia, Pneumothorax
	On the current radiograph, there are small bilateral **pleural effusions. Effusions** are better appreciated on the lateral than on the frontal radiograph. Borderline size of the cardiac silhouette without pulmonary edema persists. No pneumothorax.	a small **right apical pneumothorax** is redemonstrated. the large based left sided **pleural effusion** is in unchanged small. no appreciable pleural effusion. there is no new consolidation or pneumothorax.	as compared to the previous radiograph, there is no relevant change from the patient in the left pleural. possible small **right pleural effusion.** the cardiac silhouette is stable in size. there is no focal consolidation, effusion or pneumothorax.	Pleural Effusion, Pneumothorax, Lung Lesion --- Exudates and Transudates, Thorax, Lung, Pleural, Pneumothorax

FIGURE 9. Examples of generated chest x-ray reports with their observations predicted by MLC. Emboldened words represent abnormal findings and red words are predicted observation labels with their corresponding probabilities. Only those labels with probability greater than 0.5 are shown.

truth. From cases 2 and 3, it can be found that the observation prediction and generated reports are highly consistent. One possible reason is that the predicted observations are fused into the decoder, helping to capture these abnormalities. Another reason may derive from the accuracy discriminator (AD), which guides the model to cover these key observations.

V. CONCLUSIONS

In this paper, a novel medical report generation framework is proposed that considers both language fluency and diagnostic accuracy. From chest-radiograph images, the encoder extracts visual features and multi-type medical concepts, and then the hierarchical decoder inserts the medical concepts in sentence and word level to generate reports. More importantly, adversarial reinforcement learning (ARL) is introduced into the training procedure of medical report generation. The encoder-decoder is viewed as a generator and the reward

modules as discriminators. In training iterations, discriminators are optimized by maximum-likelihood estimation, whereas the generator is trained by reinforcement learning. Finally, the reward modules give highly accurate rewards and the generator generates better reports. In experiments, first the high performance of the proposed full model is proved by performance comparison with several classical or recently proposed models from different aspects on two large chest X-ray datasets. Ablation studies are then conducted to verify the effectiveness of the language fluency discriminator (FD) and the diagnostic accuracy discriminator (AD), followed by trade-off parameter analysis and qualitative analysis. All of the experimental results demonstrate that the proposed fully learnable ARL architecture that combines AD and FD is superior to purely traditional optimization by cross-entropy alone, or to additional RL with manually designed reward functions.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195, 2017.

[3] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," arXiv preprint arXiv:1904.02633, 2019.

[4] Y. Xue and X. Huang, "Improved disease classification in chest x-rays with transferred features from report generation," in International Conference on Information Processing in Medical Imaging. Springer, 2019, pp. 125–138.

[5] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 721–729.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.

[8] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," arXiv preprint arXiv:1701.06547, 2017.

[9] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[11] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.

[12] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 317–325.

[13] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," arXiv preprint arXiv:1511.06732, 2015.

[14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.

[15] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 457–466.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "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, 2018, pp. 9049–9058.

[20] W. Wei, L. Cheng, X. Mao, G. Zhou, and F. Zhu, "Stack-vs: Stacked visual-semantic attention for image caption generation," arXiv preprint arXiv:1909.02489, 2019.

[21] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4894–4902.

[22] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5630–5639.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[24] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[26] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[27] Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang, "Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," Journal of Cancer, vol. 10, no. 20, p. 4876, 2019.

[28] C. Sitaula and M. B. Hossain, "Attention-based vgg-16 model for covid-19 chest x-ray image classification," Applied Intelligence, pp. 1–14, 2020.

[29] A.-A. Nahid, N. Sikder, A. K. Bairagi, M. Razzaque, M. Masud, A. Z Kouzani, M. Mahmud et al., "A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network," Sensors, vol. 20, no. 12, p. 3482, 2020.

[30] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6666–6673.

[31] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 873–881.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, pp. 311–318.

[33] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81.

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111–3119.

[36] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer, 2010.

[37] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[38] 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," vol. 23, no. 2. Oxford University Press, 2015, pp. 304–310.

[39] A. E. Johnson, T. J. Pollard, S. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr: A large publicly available database of labeled chest radiographs," arXiv preprint arXiv:1901.07042, vol. 1, no. 2, 2019.

[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[42] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

DAIBING HOU received the B.S. degree in the Automation Department from Qingdao University of Science and Technology in 2018. He is currently pursuing the M.S. degree in the School of Control Science and Engineering at Shandong University.

He is the author of two articles. His research interests include pattern recognition, computer vision, and related disciplines.

ZIJIAN ZHAO received the M.S. degree in Electrical Engineering from Shandong University in 2005 and the Ph.D. degree in image processing and pattern recognition from Shanghai Jiao Tong University in 2009.

He was a Postdoctoral Researcher at the University of Oulu in Finland from 2009 to 2010. In 2010, he was a research engineer at TIMC-IMAG, University Joseph Fourier, France. In July 2012, he joined Shandong University as an Associate Professor. He is the author of more than 20 articles. His research interests include computer vision, robot vision, and computer-assisted surgery.

YUYING LIU received the B.S. degree in the Automation Department from Henan University of Technology, Zhengzhou, Henan, in 2018. She is currently pursuing the M.S. degree in Control Science and Engineering at Shandong University, Jinan, Shandong, CN.

She is the author of two articles. Her research interests include pattern recognition, computer vision, and related disciplines.

FALIANG CHANG received the B.S. degree in the Automation Department from the Shandong University of Technology, Jinan, Shandong in 1986, and the M.S. degree in Automation Department from Shandong University of Technology, Jinan, Shandong in 1989. He received the Ph.D. degree in Engineering from Shandong University, Jinan, Shandong in 2005.

He began teaching in the Department of Automation of Shandong University of Technology in 1989, was promoted to a lecturer in 1992, an associate professor in 1996, and a professor in Shandong University in 2000. In 2007, he was a visiting scholar at the State University of Michigan's School of Engineering (USA) as a visiting scholar. He is the author of more than 70 articles. His research interests include pattern recognition, computer vision, biometric recognition and authentication, and related disciplines.

He holds seven patents. He oversees the subject of pattern recognition and intelligent systems, and he is the director of the "Engineering System Control" laboratory at the Shandong Provincial Key Laboratory. Furthermore, he is a member of the Shandong Provincial Informatization Expert Group and the deputy director of the Automation Technology Committee of the Shandong Automation Institute.

SANYUAN HU graduated from Shandong Medical University in July 1987 and entered the Second Affiliated Hospital of Shandong Medical University the same year. He has successively served as resident, chief physician, deputy chief physician, and chief physician.

He was the director of surgery and general surgery of Qilu Hospital of Shandong University in 2003 and the director of endoscopic diagnosis and treatment technology training base of the Ministry of Health of Qilu Hospital of Shandong University in 2008. In 2011, he was appointed Vice President of Qilu Hospital of Shandong University. In 2012, he was hired as "Mount Tai Scholar" distinguished professor of Shandong Province. In 2019, he was appointed the president of Shandong Qianfoshan Hospital (probation period of 1 year). In the field of laparoscopic research, Professor Hu Sanyuan led the team that won the first prize for scientific and technological progress in Shandong Province. He has won nine other scientific research awards at the provincial and ministerial levels, in addition to publishing more than 30 SCI papers and applying for two invention patents. He has published 16 monographs and translated works and five audio-visual teaching materials. He is editor-in-chief of the Journal of Laparoscopic Surgery and Journal of Clinical Practical Surgery, deputy editor-in-chief of the China Journal of Endoscopy and the China Journal of Modern Medicine, and a standing or editorial board member of 20 magazines.

He was an outstanding academic leader in Shandong's health system in 2005, a young and middle-aged key scientific and technological talent, and was awarded the title of Shandong's medical technical expert in 2006. He was awarded the highest award for endoscopic medicine by the Chinese Medical Association in 2005, 2006, and 2008; the "Endoscopic Award" in 2008; the fifth "Honorary Award for Humanities Medicine" in 2008; and the honorary title of "Outstanding Graduate Instructor of Shandong University" in 2009. Professor Huang Zhiqiang, academician of the Chinese Academy of Engineering, praised him as "one of the pioneers in developing laparoscopic surgery in China."