

Few-shot Radiology Report Generation for Rare Diseases

Xing Jia

xjia18@fudan.edu.cn

Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, China

for normal sentence generation and generation module for abnormal sentence generation. Hierarchical models inspired by [24] for paragraph generation were also introduced [10], [13], [15], [17], [31], and each of them payed attention to one aspect of the problems existed in this task, e.g, semantic repetition [15], data bias [10], [31], features fusion of images form multi views [17], and background information incorporating, etc.

In medical research, according to [32], it was reported that it could waste several years for patients with rare diseases to receive the final accurate diagnosis result with the help of 8 physicians. An effective detection of rare diseases allows the patients to receive a early and timely treatment. However, few existing research works mentioned above can actually work well for rare disease detection and generate the diagnosis report from the radiology data, which will be the main task to be studied in this paper.

Two main challenges are presented on the radiology report generation for rare diseases:

- **Rare diseases low prevalence rate.** There are only quite a small number of patient instances with rare diseases in the dataset which will result in the poor performance in their detection. Fig.2. shows the number of patients of each disease class of Indiana University Chest X-ray (IU X-Ray) [1], which is a benchmark dataset studied in the task of radiology report generation [8]–[10], [12], [13], [15]–[18]. It follows the long-tail distribution, and many disease classes only have a few patient instances in the dataset.
- **Correlations among diseases.** These diseases can supply each other with complementary information so as to benefit the rare diseases detection and their corresponding report generation. A graph was proposed in [9] to model the correlations among diseases, but it was not flexible since it was designed manually by the domain expert. Also, the potential correlations exist among some diseases which are very important to rare diseases detection.

To tackle these challenges, we propose a few-shot Radiology report Generation model, namely RareGen, assembled with two components for better semantic representations learning:

- A few-shot learning generative adversarial network is introduced for generating artificial medical instances for rare diseases. We propose a Label Generator inspired by [33] for ICD coding, and cooperate it with generative adversarial network, so as to make the generated artificial instances capture the semantic information of rare diseases.
- A disease graph convolution is proposed to model and strengthen the intrinsic correlations among diseases, which allows knowledge transfer from regular diseases to those rare ones. Instead of relying on domain experts, we propose a data-driven strategy for the weight matrix construction based on the historical medical report knowledge library, which can be easily extended to other research areas. Moreover, a mechanism is introduced

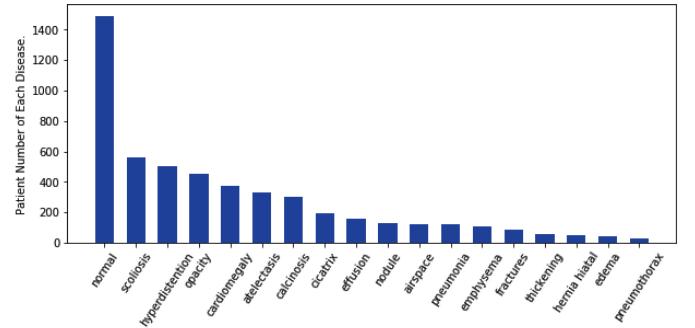


Fig. 2. The number of patients of each disease class of IU X-Ray.

which can capture the potential correlations among some diseases automatically.

The contributions are summarized as follows:

- We propose a novel radiology report generation model, i.e., RareGen, for rare diseases in this paper.
- A few-shot learning generative adversarial network is introduced for the few-shot instances augmentation, which can benefit the rare diseases detection.
- A disease graph convolution is proposed to model and strengthen the intrinsic correlations among diseases, which allows knowledge transfer from regular diseases to those rare diseases.
- A disease graph is integrated into the report generation model, which ensures the model provide more accurate reasoning.
- Extensive experiments are conducted to demonstrate the effectiveness of our model.

II. RELATED WORK

A. Image Caption and Paragraph Generation

Image caption [5] as the first attempt of image-to-text translation generally based on the encoder-decoder framework which is inspired by the recent Neural Machine Translation(NMT) [3]. The image as the source text and the corresponding caption as the target text. Generally, Convolutional Neural Network(CNN) is employed as encoder to compact the image into a lower embedding space, then Recurrent Neural Network(RNN) as decoder combined with attention mechanism accepts the visual information and outputs a sentence with variable length [6] [7]. Considering one sentence with the limited capacity of recapitulating every details in an image, the task of paragraph generation, whose goal is to depict an image in a fine-grained manner, is recently introduced [21], [24]–[26]. In a paragraph, all the sentences should keep continuity with each other, and also server the same or similar theme. While report generation task pays more attention to the abnormal or disease findings detection, especially rare diseases, and then generates a report to describe them.

B. Radiology Report Generation

Some works have explored methods for report generation based on deep learning recent years [8]–[18], for its value

in both academia and industry. Christy et al [18] proposed a graph transformer model, and it could dynamically transform image features to high-level semantics. Prior knowledge, e.g., background information or medical concepts, was incorporated to assist the model own the ability of the high-level reasoning [13] [17]. Xue et al [8] built an recurrent model to enforce the coherence between sentences.

C. Few Shot Learning

Recently, researchers have tried to bridge the gap between deep learning technique and few-shot instances. GAN-based models for few shot learning as one of potential solutions to solve this issue have been introduced in various tasks e.g., text classification [27], time series generation [28], object localization [29], attribute translation [30], etc. They are to synthesis pseudo features of few-shot cases. However, few-shot learning has not been studied in assisting report generation task so far. To the best of our knowledge, this is the first study that introduces a model for few-shot instances augmentation (i.e., generating artificial medical instances for rare diseases) for assisting the radiology report generation involves rare disease.

III. NOTATIONS, TERMINOLOGY DEFINITION AND PROBLEM FORMULATION

Prior to talking about the proposed model, we will provide the notations, terminology definitions and problem formulation in this section first.

A. Notations

In the sequel of this paper, we will use the lower case letters (e.g., x) to represent scalars, lower case bold letters (e.g., \mathbf{x}) to denote column vectors, bold-face upper case letters (e.g., \mathbf{X}) to denote matrices, and upper case calligraphic letters (e.g., \mathcal{X}) to denote sets or high-order tensors. Given a matrix \mathbf{X} , we denote $\mathbf{X}(i, :)$ and $\mathbf{X}(:, j)$ as its i_{th} row and j_{th} column, respectively. The (i_{th}, j_{th}) entry of matrix \mathbf{X} can be denoted as either $\mathbf{X}(i, j)$ or $\mathbf{X}_{i,j}$, which will be used interchangeably. We use \mathbf{X}^\top and \mathbf{x}^\top to represent the transpose of matrix \mathbf{X} and vector \mathbf{x} . For vector \mathbf{x} , we represent its L_p -norm as $\|\mathbf{x}\|_p = (\sum_i |\mathbf{x}(i)|^p)^{\frac{1}{p}}$. The Frobenius-norm of matrix \mathbf{X} is represented as $\|\mathbf{X}\|_F = (\sum_{i,j} |\mathbf{X}(i, j)|^2)^{\frac{1}{2}}$. The element-wise product and concatenation of vectors \mathbf{x} and \mathbf{y} are represented as $\mathbf{x} \odot \mathbf{y}$ and $\mathbf{x} \sqcup \mathbf{y}$. We will denote fully connected layers parameterized by variable matrix \mathbf{W} as $\text{FC}(\cdot; \mathbf{W})$; sigmoid function as $\sigma(\cdot)$; softmax function as $\text{softmax}(\cdot)$.

B. Terminology Definitions

Automatic radiology report generation aims to detect diseases from input X-ray images and generate a textual description report automatically for patients. Several important terminologies used in this paper can be defined as follows, which include *patient instance*, *rare disease* and *disease graph*.

Definition 1 (Patient Instance): Formally, we can denote the sets of patients studied in this paper as $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$. Medical data available for each patient $p_i \in \mathcal{P}$ in our dataset

can be denoted as a pair (I_i, R_i) , where I_i denotes the X-ray image and R_i denotes the textual report of p_i , respectively.

Definition 2 (Disease Set): Formally, we can represent the set of diseases studied in this paper as $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$. For each patient instance p_i in our dataset, he/she can be either healthy (without any diseases) or sick, and we can represent the health status of patient p_i by a multi-hot label vector $\mathbf{l}_i = [l_{i,1}, l_{i,2}, \dots, l_{i,m}]$, where entry $l_{i,j} \in \{0, 1\}$ denotes patient p_i has disease d_j or not. For the diseases with very few infected patients, we can name them as the rare diseases in this paper.

Definition 3 (Disease Graph): Among the diseases studied in this paper, there may exist extensive correlations, since many diseases may co-appear in many patients. To denote such disease correlations, we introduce the disease graph in this paper, which can be denoted as $G = (\mathcal{D}, \mathcal{E})$. The set \mathcal{E} denotes the edges among the diseases in \mathcal{D} . For the rare diseases, via such disease edges, information from the common diseases can be effectively transferred and utilized to improve their detection results.

In this paper, we will construct a unique disease graph for each patient, and its structure can be learned with both patient' X-ray images representation learning and historical report libraries, which will be introduced in detail in the following method section. Given the medical graph G , we can also denote its structure as the adjacency matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$, where entry $\mathbf{M}(i, j)$ denotes the correlation between diseases d_i and d_j . The concrete representation of matrix \mathbf{M} for each patient will be introduced in the method section as well.

C. Problem Formulation

Based on the above descriptions and terminology definitions, we can formulate the problem studied in this paper as follows:

Problem Formulation: Formally, given the patient set \mathcal{P} , in this paper, we aim to build a model $f : I_i \rightarrow R_i$ to project the X-ray medical image to the textual medical report for each patient $p_i \in \mathcal{P}$ in our dataset. Such X-ray medical image and textual medical report of patients are all about diseases as defined in set \mathcal{D} . Different from the existing works, we cast an extra requirements on the learned model f and the generated medical report, which should effectively detect rare diseases which don't appear frequently among patients in our dataset.

IV. METHODS

In this section, we will elaborate RareGen for rare disease report generation. As shown in Fig.3., the whole framework of RareGen can be branched into multi-label classification and report generation. While doing classification, a few-shot learning generative adversarial network and the disease graph are proposed for better semantic information learning, which will be introduced in part A and part B. The hierarchical decoder for report generation is introduced in Part C.

A. Few-shot Instance Generation

The few-shot learning generative adversarial network we proposed is to generate artificial medical instances for rare

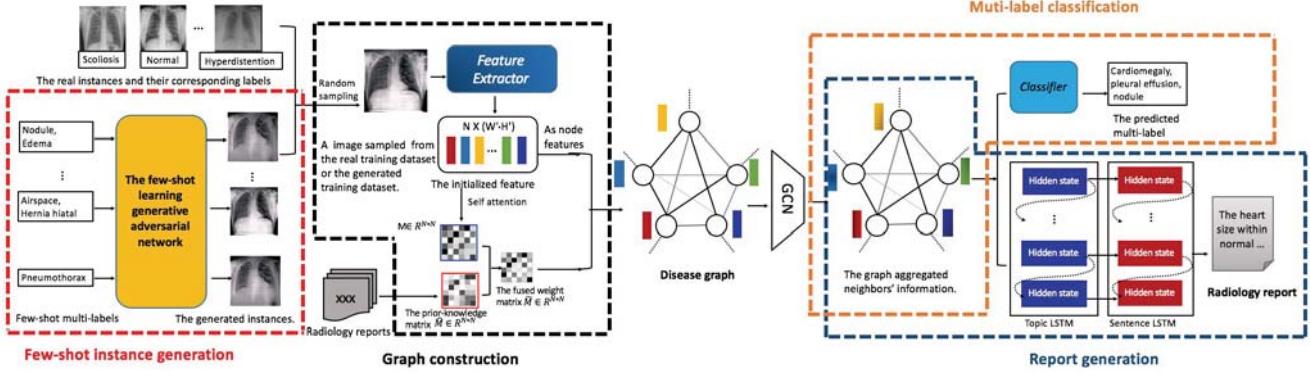


Fig. 3. The overview of our radiology report generation for rare diseases.

diseases in the latent image space, as shown in Fig.4., for better representations learning of both rare and regular diseases. To avoid the problem of model collapse while training it, WGAN-GP [22] is employed. Specifically, we transform the discrete few-shot multi-label \mathbf{l} into a real-valued vector \mathbf{l}' firstly. Then, the Generator Gen , contains several stacked de-convolution layers, takes \mathbf{l}' , the concatenation of \mathbf{l}' and a random Gaussian noise vector \mathbf{n} , as input to generate a fake image $\hat{\mathbf{I}} = Gen(\mathbf{l}', \mathbf{n})$. The discriminator Dis , contains two-stacked linear layers followed by the activation function LeakyReLU, takes either the fake-pair (the generated image $\hat{\mathbf{I}}$ and \mathbf{l}') or real-pair (the real image \mathbf{I} and \mathbf{l}') as input to produce a real-valued score $v = Dis(\hat{\mathbf{I}}/\mathbf{I}, \mathbf{l}')$ activated by $\sigma(\cdot)$, representing how realistic the pair is. Note that, we perform convolution operation on $\hat{\mathbf{I}}$ and \mathbf{I} before it concatenate with \mathbf{l}' . The optimization of this process can be defined as:

$$L_{WGAN-GP} = Dis(Gen(\mathbf{l}', \mathbf{n}), \mathbf{l}') - Dis(\mathbf{I}, \mathbf{l}') + \beta \times (|\nabla Dis(\hat{\mathbf{I}}, \mathbf{l}')| - 1)^2, \quad (1)$$

where $\tilde{\mathbf{I}} = \gamma \times \mathbf{I} + (1 - \gamma) \times \hat{\mathbf{I}}$, and $\gamma \in (0, 1)$ is a hyper parameter. $\beta \times (|\nabla Dis(\hat{\mathbf{I}}, \mathbf{l}')| - 1)^2$ is the penalty term.

To further ensure the generated instance capture the semantic information of its input multi-label \mathbf{l} , we further design a Label Generator, contains several stacked convolution layers, to reconstruct the label \mathbf{l} from the generated instance $\hat{\mathbf{I}}$,

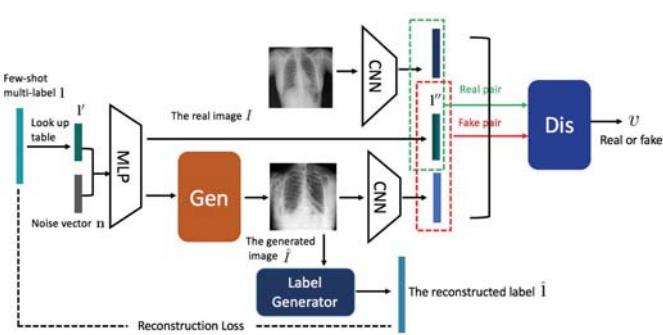


Fig. 4. The diagram of the few-shot learning generative adversarial network.

obtaining the reconstructed label $\hat{\mathbf{l}}$. The reconstruction loss is defined as:

$$L_{recons} = -\text{cosine_sim}(\hat{\mathbf{l}}, \mathbf{l}), \quad (2)$$

where $\text{cosine_sim}(\cdot, \cdot)$ is the cosine similarity function.

B. Disease Graph Construction

(1) Weight Matrix Construction

Graph Convolution Network (GCN) has been confirmed to be effective to learn more informative representations considering both node features and structure information among them [23]. It is employed to enhance the semantic representations of both rare and regular diseases. This process can be formally defined as:

$$\mathbf{H}^{(k+1)} = \text{ReLU}(\mathbf{M}\mathbf{H}^k\mathbf{W}^k), \quad (3)$$

where $\mathbf{M} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \in R^{N \times N}$ is adjacency matrix defined by prior knowledge, and $\mathbf{D} \in R^{N \times N}$ is the corresponding diagonal matrix of \mathbf{A} . For the construction of prior-knowledge matrix \mathbf{M} , the entry $\mathbf{M}(i, j)$ is set to 1 if disease d_i and disease d_j appear in the same sentence of a report; otherwise 0. In this way, the weight matrix constructed in [9] can be enhanced. $\mathbf{H}^{(k)} \in R^{N \times d}$ and $\mathbf{W}^{(k)} \in R^{d \times d'}$ are the node representations matrix and the trainable linear transformation matrix of the k th layer, respectively. Here, N , d , and d' denote the number of diseases, feature dimension, and feature dimension after projecting.

To both incorporate the prior knowledge, and explicitly explore the potential correlations among diseases, we made some modification of Eq(3) by breaking the weight-matrix into two parts, one constructed by prior knowledge and the other learned by our model automatically:

$$\mathbf{H}^{(k+1)} = \phi(\tilde{\mathbf{M}}\mathbf{H}^k\mathbf{W}^k), \quad (4)$$

$$\tilde{\mathbf{M}} = \lambda\mathbf{M} + (1 - \lambda)\hat{\mathbf{M}}, \quad (5)$$

where $\hat{\mathbf{M}} \in R^{N \times N}$ is the matrix learned by our model automatically, and $\lambda \in (0, 1)$ is a hyper-parameter to control the proportion of prior weights and learned weights.

In this paper, we employ self-attention [4] to dynamically capture the correlations among diseases no matter whether they are regular or rare, and the weight matrix $\hat{\mathbf{M}}$ is the multiplication of the node features matrix and its corresponding transposition, followed by softmax along the column dimension. The formulation can be defined as:

$$\hat{\mathbf{M}} = \mathbf{H}^{(0)} \odot \left(\mathbf{H}^{(0)} \right)^\top, \quad (6)$$

$$\hat{\mathbf{M}}(i,:) = \frac{\exp(\hat{\mathbf{M}}_{ij})}{\sum_j \exp(\hat{\mathbf{M}}_{ij})}, \quad (7)$$

where $\mathbf{H}^{(0)} \in R^{N \times d}$, and each row corresponds to a initialized feature of a node in the graph. $(\cdot)^\top$ is the transpose operation. The detailed information about obtaining $\mathbf{H}^{(0)}$ is introduced in subsection (2).

(2) Node Feature Initialization

Besides the weight matrix, the other fundamental component of the disease graph convolution is node features. High-quality initialization of each node can bring a lot of benefits to the downstream task, e.g., classification and report generation in this paper. As the original features of a image is 3-dimensional tensor, to effectively extract feature corresponding to a specific disease node plays a core role during node initialization process. Inspired by the convolution-deconvolution framework proposed in [21], we propose a strategy for node features initialization named Feature Extractor (FE) whose main idea is to compress node feature into a low-dimensional vector in the manner of convolution-deconvolution.

Given the feature maps $\mathcal{V}^{conv} \in R^{1024 \times W \cdot H}$ extracted by a CNN(e.g., densenet121 [2] in this paper), feature maps are taken as the material to further process the node initialized features. Specifically, a convolutional layer acts upon \mathcal{V}^{conv} obtaining N node features $\mathcal{F}_0^{nodes} \in R^{N \times W' \times H'}$ firstly, which is applied with convolutions.

$$\mathcal{V}^{conv} = CNN(I) \in R^{1024 \times W \times H}, \quad (8)$$

$$\mathcal{F}_0^{nodes} = Conv(\mathcal{V}^{conv}) \in R^{N \times W' \times H'}, \quad (9)$$

where $Conv$ is the convolutional layer. 1024, W and H are the number of channel, width and height of the feature maps the output of densenet121 after block 4. N , W' and H' are the number of channel, width and height of feature maps after $Conv$ operation. The number of its kernels is equal to the number of nodes N in the graph, since each kernel pays a attention to a specific node visual information in the graph.

Then, a de-convolutional¹ layer is leveraged to reconstruct \mathcal{F}_0^{nodes} to the original feature maps, and smoothed $L1$ loss is employed to bound the discrepancy between the reconstructed feature maps \mathcal{V}^{deconv} and the original feature maps \mathcal{V}^{conv} at pixel-level.

$$\mathcal{V}^{deconv} = Deconv(\mathcal{F}_0^{nodes}) \in R^{1024 \times W \times H}, \quad (10)$$

¹The more accurate name should be transpose convolution. We use deconvolution in this paper for simplicity.

$$L_node = L_{1norm}(\mathcal{V}^{conv}, \mathcal{V}^{deconv}), \quad (11)$$

where $Deconv$ is a de-convolutional layer, and its kernel size is same to the convolutional layer $Conv$. N , W' and H' are the number of channel, width and height of feature maps after $Deconv$ operation.

In this way, each node initialized feature is enforced to own representative information. \mathcal{F}_0^{nodes} is further reshaped to $\mathcal{F}_0^{nodes'} \in R^{N \times (W' \cdot H')}$, and each vector along column dimension represents one node initialized feature.

C. Report Generation

Inspired by the hierarchical LSTM structure applied in paragraph generation task [21] [24] [26], it is employed as decoder for report generation. Specifically, the hierarchical LSTM consists of a topic $LSTM^T$, responsible for topic generation, and a sentence $LSTM^S$, responsible for its corresponding sentence generation, as illustrated in Fig.3.

At each time step t for generating the topic vector \mathbf{h}_t^T , the input vector \mathbf{x}_t^T of $LSTM^T$ is defined as the concatenation of context vector \mathbf{c}_t^T and \mathbf{h}_{t-1}^T , since \mathbf{c}_t^T can supply the contextual information of a sentence to be generated, and \mathbf{h}_{t-1}^T ensures the dependency of two sentences at the same time. The initialization of topic LSTM \mathbf{h}_0^T and \mathbf{c}_0^T are mean-pooled image features of densenet121 after block 4 projected by two-linear layers.

$$\mathbf{h}_t^T = LSTM^T(\mathbf{x}_t^T, (\mathbf{h}_{t-1}^T, \mathbf{c}_{t-1}^T)), \quad (12)$$

$$\mathbf{x}_t^T = \mathbf{W}_c(\mathbf{c}_t^T \sqcup \mathbf{h}_{t-1}^T), \quad (13)$$

where \mathbf{W}_c is the transformation matrix, and \sqcup is the concatenation operation.

To dynamically capture the contextual information, attention mechanism [6] is applied. Given the hidden state \mathbf{h}_{t-1}^T of $LSTM^T$, the normalized attention distribution over all nodes features $\mathcal{F}^{nodes'}$ is obtained. The context vector \mathbf{c}_t^T is thus calculated by aggregating all nodes features weighted by attention.

$$a_{(N,t-1)} = \mathbf{W}_{att}[\tanh(\mathbf{W}_h \mathbf{h}_{t-1}^T + \mathbf{W}_n \mathcal{F}^{nodes'})], \quad (14)$$

$$\alpha_{(N,t-1)} = softmax(a_{(N,t-1)}), \quad (15)$$

$$\mathbf{c}_t^T = \sum_{n=1}^N \alpha_{(n,t-1)} \mathcal{F}_n^{nodes'}, \quad (16)$$

where \mathbf{W}_h , \mathbf{W}_n and \mathbf{W}_{att} are transformation matrix for the hidden state of $LSTM^T$ at time step $t-1$, node features matrix $\mathcal{F}^{nodes'}$, and parameters of attention network, respectively.

Once the topic vector \mathbf{h}_t^T is obtained, it is taken as the first input of $LSTM^S$, which explicitly force the sentence to be generated server the topic \mathbf{h}_t^T . The subsequent inputs of $LSTM^S$ are its generated words' embedding \mathbf{e}_{t-1}^S . The hidden state \mathbf{h}_t^S of $LSTM^S$ is further projected to vocabulary-sized vector normalized by softmax, and the corresponding word with maximum probability is selected.

$$\mathbf{h}_t^S = LSTM^S(h_t^T / \mathbf{e}_{t-1}^S, (\mathbf{h}_{t-1}^S, \mathbf{c}_{t-1}^S)), \quad (17)$$

$$\mathbf{w}_t^S = \text{argmax}(\text{softmax}(\mathbf{W}_w \mathbf{h}_t^S)), \quad (18)$$

where \mathbf{W}_w is transformation matrix of a linear layer, and argmax is index selection operation.

Note that topic vector is also acted as a indicator to determine the number of generated sentences. Specifically, each topic vector is projected to a distribution over two states {CONTINUE=0, STOP=1}, which determines whether the sentence is the last one in a report R_i .

V. EXPERIMENTS AND RESULTS

A. Dataset

We adopt the public dataset IU X-Ray [1]. It contains 3955 radiology reports, and each report is associated with two images of different views.

For data preprocessing, we tokenize all the words in the impression and findings sections, convert them to lower-cases, and eliminate tokens by minimum frequency 3 obtaining a vocabulary with 1,108 unique words. We use a special token $\langle unk \rangle$ to represent the eliminated tokens. Each sentence is added tokens $\langle start \rangle$ and $\langle end \rangle$ indicating the start and end of a sentence/sequence (e.g., the baseline [5]–[8] take the whole report contains several sentences as one sequence). The sentence/sequence whose length is less than the pre-defined maximum length of a sentence/sequence is padded with the special token $\langle pad \rangle$.

To evaluate our model, we adopt a commonly used splitting method following [9], [10], [12], [13], [16]–[18]: we randomly split dataset into training, validation and testing by a ratio, i.e., 8:1:1 in this paper. There is no overlap among different sets.

B. Experiment Settings

We pre-train RareGen on cheXpert [19] a public chest X-Ray dataset, and then train it on IU X-Ray for 64 epochs with loss function Binary Cross Entropy (BCE). Adam is used for training, with a batch size of 16, a weight decay of 1e-5, and the initial learning rate of 1e-4. Scheduler MultiStepLR is employed with epoch range [15, 30, 50] and gamma 0.1. Each image is center-cropped to 512×512 without data augmentation. We extract visual features after the block 4 of densenet121 [2], yielding $1024 \times 16 \times 16$ feature maps. For stable training of the few-shot learning generative adversarial network, batch normalization is adopted. For decoder, the topic LSTM and sentence LSTM are single-layer LSTMs with hidden dimension of 512. Each input word is encoded as a embedding with the dimension of 256. We set the maximum number of sentences of a report and maximum number of tokens in a sentence as 10 and 30. Cross Entropy is employed as the loss function. All the implementations are based on PyTorch and conducted on NVIDIA Tesla V100.

C. Metrics

For evaluating report generation, we follow most studies and adopt the metrics originally developed for the evaluation of machine translation, or text summaries, including BLEU1, 2, 3, 4, CIDEr, ROUGE-L [8]–[18]. Also, there are some

metrics proposed for report generation, and we adopt KA [8] as it has been used in several studies [8], [10]. Specifically, we extract key words in the MTI annotations of IU X-Ray dataset to construct a key-word dictionary containing 421 words. During inference, KA metric is the ratio of the number of keywords correctly generated by the model to the number of all keywords in the ground-truth.

However, these metrics are purely based on words occurrences and thus fail to examine whether the generated reports have the correct semantic meaning. For example, the reference sentence is "there are degenerative changes of the spine.", and sentences generated by RareGen and CNN_RNN are "degenerative changes are present in the spine." and "there are **no** degenerative changes of the spine.", respectively. High BLEU score can be obtained of the pair "there are degenerative changes of the spine." and "there are **no** degenerative changes of the spine.", while the meanings they expressed are totally different. Meanwhile, report generation performance of classification based models depends on the classification results to large extent. Still, the meanings of these two sentences can be demonstrated by the classification results of a certain class among the labels. To further measure the quality of report generation, we study the multi-label classification results of classification. So we use Accuracy, F_1, Precision, Recall, and Area Under the ROC Curve (AUC) to further evaluate the performance of disease classification, and micro-average is adopted (AUC, Precision, Recall and F_1).

D. Baselines

(1)Baselines for report generation

We compare our full model RareGen with the methods: Feedback [8], A3FN [10], CNN-RNN [5], Soft-Att [6] and Att-RK [7], Co-Att [12] and KG [9]². CNN-RNN, Soft-Att and Att-RK are state-of-the-art image caption models, and others are radiology report generation models.

(2)Baselines for classification

Since performance of classification based models depend on the classification results to large extent, we further classify the baselines into: 1). classification based models: Co-Att, A3FN, KG and RareGen. 2). non-classification based models: Feedback, CNN-RNN, Soft-Att and Att-RK. We compare the classification results of RareGen with classification based models, some classical Deep CNN models applied in ChestXray8 [20] and also densenet121 [2] the backbone in this paper.

E. Experimental Results and Analysis

(1)Results on report generation

Table I shows results of automatic evaluation comparing RareGen with the baselines introduced before. Most importantly, RareGen outperforms all baseline models by great margins on KA metric (e.g. 17.6(RareGen) VS 7.5(CNN_RNN)), which demonstrates that RareGen owns a stronger ability to distill disease features by the disease graph convolution we proposed. RareGen achieves slightly lower BLEU_1, 4 score

²Note that both the codes of Co-Att and Feedback are not released, and we re-implement both of them.

Two-view images				
Ground truth	right basilar airspace disease, the heart is normal in size , the pulmonary vascularity is within normal limits in appearance . <code><u>cm nodule within the lung base seen only on the lateral view, no pneumothorax or pleural effusion , patchy right lower lung opacification is noted .</code>	cardiomegaly and increased interstitial opacities xxxx represent interstitial edema . cardiomegaly , mediastinal contours are normal limits . increased interstitial opacities , no pneumothorax or large pleural effusion , no acute osseous abnormality .	xxxx xxxx opacity in the left base xxxx interelectasis , heart size moderately enlarged stable mediastinal contours . xxxx xxxx opacity in the left lung base , otherwise no focal alveolar consolidation no definite pleural effusion seen , no typical findings of pulmonary edema .	
Co-Att[12]	cardiomedastinal silhouette is normal in size and contour, right lung volumes are clear, no pneumothorax or large pleural effusion	stable cardiomegaly, stable cardiomegaly, no focal consolidation, mediastinal contours are normal limits, increased interstitial opacities, no pneumothorax or large pleural effusion, no acute osseous abnormality	heart size and mediastinal contour are within normal limits, lungs are hyperexpanded with flattened diaphragms of the diaphragm, lungs are clear without focal airspace disease, no acute bony abnormality	
KG[9]	no acute cardiopulmonary abnormality , no focal consolidation , no pneumothorax or pleural effusion heart size is within normal limits . calcified right hilar lymph xxxx . there is a calcified granuloma in the left lung base . no focal consolidation .	cardiomegaly with mild interstitial edema , there is a left basilar airspace opacity , there is a left basilar opacity , there is a left basilar opacity , there is a left basilar opacity . there is a left basilar opacity .	no acute cardiopulmonary abnormality , the lungs are clear bilaterally , specifically no evidence of focal consolidation pneumothorax or pleural effusion . cardio mediastinal silhouette is unremarkable . visualized osseous structures of the thorax are without acute abnormality .	
RareGen	no acute cardiopulmonary abnormality , there is a right lower lobe airspace disease , there is no pneumothorax or large pleural effusion, the heart is normal in size , the mediastinum is unremarkable . there is a <code><u>cm right upper lobe nodule .</code>	cardiomegaly with mild interstitial edema , no focal consolidation pneumothorax or large pleural effusion , negative for acute bone abnormality .	cardiomegaly with mild bibasilar atelectasis , no focal airspace consolidation , no pleural effusion or pneumothorax , heart size is enlarged , increased interstitial opacities .	

Fig. 5. 3 cases with two-view of chest X images. The ground truth and generated reports by RareGen, Co-Att and KG are shown.. The sentences marked in blue are the description of rare diseases or disease with a few positive instances in the ground truth. The sentences marked in red are the precise corresponding description generated by RareGen or baselines.

than Co-Att and A3FN. But, it doesn't mean that RareGen in a worse performance, since high BLUE score does not necessarily imply the correctness of a generated sentence as we explained before. Besides, KG, A3FN, Co-Att and RareGen surpass CNN_RNN, Soft_Att, Att-RK and Feedback by relatively large margins on almost all metrics, which indicates classification based models are beneficial to generating structured reports comparing with non-classification based models possibly due to the better representations that classification based models learned.

TABLE I
PERFORMANCE COMPARISON OF REPORT GENERATION
ON IU X-RAY DATASET

Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4
CNN_RNN [5]	0.295	0.216	0.158	0.112
Soft_Att [6]	0.363	0.257	0.183	0.135
Att-RK [7]	0.344	0.251	0.168	0.116
Feedback [8]	0.434	0.331	0.234	0.177
Co-Att [12]*	0.455	0.288	0.205	0.154
A3FN [10]*	0.443	0.337	0.236	0.181
KG [9]*	0.441	0.291	0.203	0.147
RareGen	0.448	0.343	0.255	0.178
Methods	CIDEr	ROUGE-L	KA(%)	
CNN_RNN [5]	0.136	0.258	7.5	
Soft_Att [6]	0.288	0.342	8.3	
Att-RK [7]	0.192	0.358	7.8	
Feedback [8]	0.312	0.351	10.1	
Co-Att [12]*	0.277	0.369	11.37	
A3FN [10]*	0.374	0.347	13.5	
KG [9]*	0.304	0.367	15.8	
RareGen	0.378	0.371	17.6	

Note that *score is taken from its corresponding paper, except for KA metric, so the data split can be different from RareGen.

(2)Results on multi-label classification

The multi-label classification results are shown in Table II. RareGen outperforms all baselines of disease classification on Accuracy, F_1 and Precision, which demonstrates its effectiveness and also implicitly interprets the high KA score our

model obtained. RareGen achieves over 6% higher in accuracy comparing with KG, demonstrating RareGen's stronger ability of diseases detection. We can also find that all models are with low accuracy scores. The reason can be we impose a strong constrain on accuracy metric, and a instance is considered to be hitted only if all the labels of the instance are predicted correctly. While it is meaningful to study accuracy metric since it indicates whether a report can be generated completely and correctly.

TABLE II
PERFORMANCE COMPARISON OF CLASSIFICATION ON IU X-RAY DATASET.

Methods	Accuracy	F_1	Precision	Recall	AUC
ChestXray8 [20]	0.719*	-	-	-	-
DenseNet121 [2]	0.2240	0.2617	0.1617	0.6861	0.7831
TieNet [11]	0.2283	0.3259	0.2643	0.3987	0.7866
Co-Att [12]	0.2427	0.3361	0.2891	0.4014	0.7909
KG [9]	0.2205	0.3740	0.3203	0.4492	0.8222
RareGen	0.2889	0.3798	0.4580	0.3244	0.8136

(3)Ablation study

To verify the effectiveness of the components(the few-shot learning generative adversarial network and FE), we further do some ablation study. We compare our full mode RareGen trained on the fused dataset(the original train dataset and the artificial dataset generated by the few-shot learning generative adversarial network) with FE, with the variants RareGen(-), trained on the fused dataset without FE, and also RareGen(-, -) solely trained on the original dataset without FE. The classification and report generation results are shown in Table III and Table IV. On the whole, we can observe that both FE and the few-shot learning generative adversarial network make contributions to the performance improvement to the task of classification and report generation. Most notably, RareGen(-) increases accuracy score by near 3% and KA score by over 3% compared to RareGen(--), demonstrating the crucial

of the few-shot learning generative adversarial network in solving the issue of few-shot learning on report generation task. Meanwhile, RareGen achieves higher automatic evaluation scores compared with RareGen(-), which confirms the effectiveness of FE. Fig.5. shows the qualitative results of RareGen and some baselines. It can be observed that the rare diseases "airspace", "nodule", and "opacities" are accurately detected by RareGen, while the Co-Att and KG can hardly detect them, which demonstrates the effectiveness of RareGen for rare diseases report generation.

TABLE III

PERFORMANCE COMPARISON OF CLASSIFICATION ON IU X-RAY DATASET

Methods	Accuracy	F_1	Precision	Recall	AUC
RareGen(-, -)	0.2462	0.3437	0.4334	0.2847	0.8084
RareGen(-)	0.2715	0.3563	0.4695	0.2870	0.8036
RareGen	0.2889	0.3798	0.4580	0.3244	0.8136

TABLE IV

PERFORMANCE COMPARISON OF REPORT GENERATION ON IU X-RAY DATASET

Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4
RareGen(-, -)	0.421	0.310	0.209	0.158
RareGen(-)	0.433	0.329	0.225	0.135
RareGen	0.448	0.343	0.231	0.178
Methods	CIDEr	ROUGE-L	KA(%)	
RareGen(-, -)	0.334	0.346	12.5	
RareGen(-)	0.383	0.366	15.9	
RareGen	0.378	0.371	17.6	

VI. CONCLUSION

In this paper, we propose a few-shot radiology report generation model RareGen for rare disease report generation. RareGen obtains a better performance for rare diseases report generation comparing with the stat-of-the-art baselines. Extensive experiments are conducted demonstrating the effectiveness of RareGen. And it can also be applied to other disease datasets, e.g., tongue images dataset [34].

VII. ACKNOWLEDGEMENT

This work is funded in part by the Shanghai Science and Technology Development Fund No. 19511121204, No.19DZ1200802 and the National Natural Science Foundation of China Projects No. U1636207. This work is also partially supported by NSF through grant IIS-1763365 and by FSU.

REFERENCES

- [1] Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA* 23, 304310 (2015)
- [2] Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 47004708.
- [3] Bahdanau D , Cho K , Bengio Y . Neural Machine Translation by Jointly Learning to Align and Translate[J]. *Computer ence*, 2014.
- [4] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. *arXiv*, 2017.
- [5] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]// In *CVPR*. 2015: 3156-3164.
- [6] 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.
- [7] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: *Proceedings of CVPR* (2016)
- [8] Xue Y, Xu T, Long L R, et al. Multimodal recurrent model with attention for automated radiology report generation[C]//In *MICCAI*, 2018: 457-466.
- [9] Zhang Y, Wang X, Xu Z, et al. When Radiology Report Generation Meets Knowledge Graph[J]. 2020.
- [10] Xie X, Xiong Y, Philip S Y, et al. Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation[C]//In *DASFAA*, 2019: 448-452.
- [11] Wang X, Peng Y, Lu L, et al. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays[C]//In *CVPR*. 2018: 9049-9058.
- [12] Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports[J]. In *ACL*, 2017.
- [13] Huang X , Yan F , Xu W , et al. Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation[J]. *IEEE Access*, 2019, PP(99):1-1.
- [14] 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, 2019.
- [15] Yin C , Qian B , Wei J , et al. Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network[C]// In *ICDM*, 2020.
- [16] Li C Y , Liang X , Hu Z , et al. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation[J]. 2018.
- [17] Yuan J , Liao H , Luo R , et al. Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment[J]. 2019.
- [18] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019.
- [19] Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoor, B.; Ball, R.; Shpan- skaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison.
- [20] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*.
- [21] Wang J, Pan Y, Yao T, et al. Convolutional auto-encoding of sentence topics for image paragraph generation[J].2019.
- [22] Gulrajani I , Ahmed F , Arjovsky M , et al. Improved Training of Wasserstein GANs[J]. 2017.
- [23] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 1, 2, 3
- [24] Jonathan Krause, Justin Johnson, Ran- jay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 2017.
- [25] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *ICCV*, 2017.
- [26] Moitreyra Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *ECCV*, 2018.
- [27] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.
- [28] Smith K E , Smith A O . Conditional GAN for timeseries generation[J]. 2020.
- [29] Lin J , He X . Few-shot Learning with Weakly-supervised Object Localization[J]. 2020.
- [30] Durall R , Pfreundt F J , Keuper J . Semi Few-Shot Attribute Translation[J]. 2019.
- [31] Harzig P , Chen Y Y , Chen F , et al. Addressing Data Bias Problems for Chest X-ray Image Report Generation[J]. 2019.
- [32] Shire, H. 2013. Rare disease impact report:Insights from patients and the medical community
- [33] Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, Eric P. Xing: Generalized Zero-Shot Text Classification for ICD Coding. In *IJCAI* 2020
- [34] A, Dan Shi , et al. "An annotated dataset of tongue images supporting geriatric disease diagnosis." In *Data in Brief* 2020.