

Received September 9, 2019, accepted October 10, 2019, date of publication October 14, 2019, date of current version November 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947134

Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation

XIN HUANG^{1,2}, FENGQI YAN¹, WEI XU³, AND MAOZHEN LI⁴

¹Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

²Software College, Jiangxi Agricultural University, Nanchang 3300029, China

³Department of Ophthalmology, Tongji Hospital, Tongji University School of Medicine, Shanghai 200065, China

⁴Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, U.K.

Corresponding author: Fengqi Yan (yanfengqi@heyeah.cn)

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 16511102800, and in part by the Fundamental Research Funds for the Central Universities under Grant 22120180117.

ABSTRACT Chest X-ray images are widely used in clinical practice such as diagnosis and treatment. The automatic radiology report generation system can effectively reduce the rate of misdiagnosis and missed diagnosis. Previous studies were focused on the long text generation problem of image paragraph, ignoring the characteristics of the image and the auxiliary role of patient background information for diagnosis. In this paper, we propose a new hierarchical model with multi-attention considering the background information. The multi-attention mechanism can focus on the image's channel and spatial information simultaneously, and map it to the sentence topic. The patient's background information will be encoded by the neural network first, then it will be aggregated into a vector representation by a multi-layer perception and added to the pre-trained vanilla word embedding, which finally forms a new word embedding after fusion. Our experimental results demonstrated that the model outperforms all baselines, achieving the state-of-the-art performance in terms of accuracy.

INDEX TERMS Attention mechanism, deep learning, radiology report generation, word embedding.

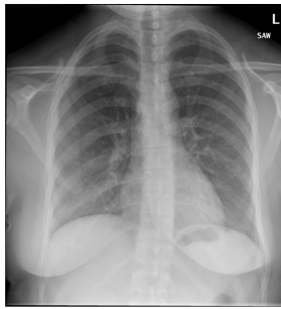
I. INTRODUCTION

Chest X-ray is a kind of medical image, which is widely used to diagnose and treat pneumonia, pneumothorax and other diseases. The diagnostic report serves as an interpretation of the medical image and describes the findings of each body region in the image, especially whether each region is normal, abnormal or potentially abnormal. Figure 1 is an example of a radiology report, in which the "finding" section lists the radiology observations for each area, and the "impression" section is the diagnosis provided by radiologists. Obviously, the report contains a lot of medical terms, writing a report requires a professional radiologist to be competent, which is time-consuming. Due to the lack of professional radiologists, the proportion of misdiagnosis and missed diagnosis is higher in regions with backward medical treatment [1]. A computer-aided radiology report generation system can lighten the

workload for radiologists considerably and assist them in decision making.

It is a complicated problem to generate a report based on radiology images. First, it is necessary to accurately find the abnormal part of the image [2]–[6], and then describe it in the form of text. Most existing literature pertaining to radiology report generation problems is based on deep learning techniques, following the encoder-decoder architecture of the machine translation task. Overall, there are two problems with the current research. First, the research focuses on how to improve the performance of decoders, for example, combining retrieval and generation for the template characteristics of the report [7]. For image decoders, more mature convolutional neural networks (CNN) such as ResNet [8] and DenseNet [9] are used to extract features. The extracted features are either directly input to the decoder, or combined with the semantics input decoder using the spatial attention mechanism [10]. Compared with taking a certain layer of features of CNN as the input of the decoder, the attention

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.



Indication : Shortness of breath
Finding : Cardiomegaly is present. The pulmonary vascularity appears within normal limits. Pacemaker is noted. Vascular calcification is seen. The lungs are free of focal airspace disease. No pneumothorax or pleural effusion is noted. Degenerative changes are present in the spine.
Impression : Cardiomegaly without overt heart failure

FIGURE 1. An example of a chest X-ray report. In the indication section lists the abnormal physical symptoms described by the patient, which is background information. In the finding section lists radiological observations for each area of body. In the impression section, the radiologist provides the diagnosis.

mechanism can enhance the representation of specific areas. However, in the existing literature of report generation task, spatial attention of image is mostly considered, while channel attention is not. Woo *et al.* [11] have shown that increasing channel attention in the model can significantly improve the performance of the extracted image features. In the task of image caption, SCA-CNN [12] also proves that the image features simultaneously perform spatial attention and channel attention operation, which can better guide the decoder to generate sentences. Second, the influence of background information related to diagnosis on report generation has been neglected in the existing studies. As shown in Fig. 1, the indication section displays background information that includes information about the patient's gender, past medical history and Abnormal physical symptoms described by the patient. The background information of the observed objects is also one of the important factors in the auxiliary diagnosis of the radiologist based on medical images. In fact, the abnormal physical symptoms described by patients are the prerequisite for the diagnosis of radiologists, which can effectively help radiologists judge according to medical images and eliminate some unnecessary interference, thus improving the diagnostic accuracy of radiologists. In the field of computer-aided diagnosis, it has been shown that the background information of patients has a positive effect on assisted diagnosis. For example, Baltruschat *et al.* [13] added patients' age, gender and other non-image information in the study of the classification of diseases on chest radiographs, which significantly improved the efficiency of classification of diseases. Zhang *et al.* [14] in the task of "finding" to generate "impressions", incorporating patient background information into the encoder can improve the BLEU-4 value of the generated sentence.

In this work, we designed a novel automatic chest X-ray radiology report generation model to solve the above problems. Our model uses a multi-attention mechanism that includes channel attention and spatial attention to enhance the mapping of sentence topics to image feature representations. Channel attention focuses on the entities in the image (what is it), and spatial attention focuses on the location of

the entities (where is it). For the decoder, the hierarchical RNN method has proven to be effective in image paragraphs generation, and we follow this approach. At the same time, we incorporate the patient's background information into the original word embedding. After the background information is encoded by the neural network, it is aggregated into a vector representation by a multi-layer perceptron and added to the pre-trained common word embedding, and finally, a new word embedding after fusion is formed.

Overall, the main contributions of our work are:

- We propose a multi-attention mechanism to enhance the mapping of sentence topics to image feature representations in image paragraph tasks.
- We propose a word embedding model that incorporates patient background information, which combines BiLSTM and attention to enhance the fusion of vanilla word embedding and background information.

The rest of this paper is organized as follows. Section II reviews the related work for image captioning and radiology report generation. Section III details the design of the proposed model. Section IV and V present and discuss the experimental settings and results, respectively. Finally, we draw conclusion in Section VI.

II. RELATED WORK

A. IMAGE CAPTIONING

Image Captioning is a task that automatically generates text descriptions for given images. This task has attracted wide attention with the success of Show and Tell [15], and its follow-up Show, Attend and Tell [16]. Most of the recent studies are based on the CNN-RNN structure and strengthen the association between pictures and words through different attention mechanisms [17]–[23]. However, the sentence sequence generated by Image Captioning is usually short and describes the most prominent visual events, which cannot fully represent the rich feature information of the image. Krause *et al.* [24] designed a hierarchical recursive neural network: sentence RNN and word RNN generate paragraph descriptions of pictures, where sentence RNN determines the number of sentences of paragraphs and generates topic vectors for each sentence, and word RNN generates words of individual sentences based on this topic vector. Recurrent Topic-Transition Generative Adversarial Network (RTT-GAN) [7] builds an adversarial framework between a structured paragraph generator and multi-level paragraph discriminators. The paragraph generator generates sentences recurrently by incorporating region-based visual and language attention mechanisms at each step. Chatterjee and Schwing [25] uses "coherence vectors", "global topic vectors", and the inherent blurriness of the variational autoencoder to correlate paragraphs with images to increase paragraph generation techniques. Experiments show that outperforming existing state-of-the-art techniques on the open dataset. Melas-Kyriazi *et al.* [26] research shows that more sequence-rich paragraphs can be generated through sequence-level training combined with triple-combination

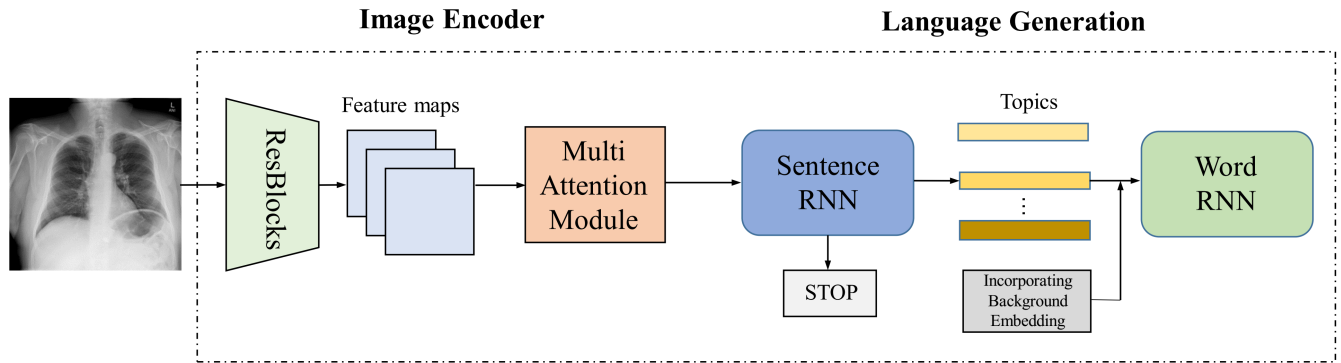


FIGURE 2. The architectural of our model consists of four sub-modules, a multi-attention module, a sentence RNN module, a background information fusion module, and a word RNN module.

penalties without changing the model architecture. This simple training method improvement has greatly improved the performance of the Visual Genome paragraph captioning dataset, with the CIDEr value increased from 16.9 to 30.6.

B. RADIOLOGY REPORT GENERATION

In recent years, many works have explored the task of radiology report generation. TieNet [27] is a chest X-ray report generation model based on the CNN-RNN architecture with attention mechanism. Jing *et al.* [10] adopted co-attention to better combine features of pictures and labels. According to the Template characteristics of the report, Li *et al.* [28] adopted a hybrid method of retrieval and reinforcement learning. For the topic generated by sentence decoder, Template Database or Generation Module was used for reinforcement learning and training, and the BLEU-4 value reached 0.15. The model designed by Xue *et al.* [29] takes into account doctors' habit of writing reports. Firstly, the diagnosis part of the report is generated, and then a complete report is generated based on the semantic features of the previous sentence and the Attention mechanism. KERP [30] breaks down medical report generation into two parts: medical anomaly graph learning and natural language modeling, dynamically transforming high-level semantics between graphical structured data (such as knowledge graphs, images, and sequences) of multiple domains, which is currently the most effective method. The performance evaluation indexes of the above work are all traditional natural language evaluation indexes. Liu *et al.* [31] pointed out that medical abnormal words and negative words should be fully considered in the evaluation process, and proposed to use Clinical Finding Scores, which can better evaluate the quality of reports.

III. METHODS

A. OVERVIEW

We propose a multi-attention hierarchical model, as shown in Fig. 2. Fundamentally, the model is an encoder-decoder architecture. The decoder receives a chest image as input and generates the feature representation of the image through a convolutional neural network and multi-attention module,

which simultaneously pays attention to the channel information (what is it) and spatial information (where is it) of the image. Next, the decoding process starts with the feature representation of the image, which is used to generate the complete report. Descriptive reports of medical images often contain paragraphs of text with multiple sentences, each focusing on a specific topic (region, representation, disease, etc.). The decoder generates long paragraph text using a hierarchical LSTM structure containing the sentence LSTM and the word LSTM. The decoder first generates a series of high-level topic vectors representing the sentences, and then generates a sentence based on each topic vector. Specifically, the feature representation of the image is input into a sentence LSTM and generates a topic vector. The topic vector represents the semantic representation of the sentence to be generated. The word LSTM takes a given topic vector and the word embedded in the background information as input, generating a series of words to form a sentence.

B. IMAGE ENCODER

The main body of the image decoder is ResNet. In order to make the sentence theme more precise focus on the corresponding image features, we added a multi-attention module containing channel attention and space attention to the original ResNet. The multi-attention module architecture is shown in Fig. 3. Suppose we want to generate the t -th topic in the report, given an intermediate feature map $V^l \in \mathbb{R}^{C \times H \times W}$ as input, where $V^l = CNN(V^{l-1})$, and we have the last hidden state of Sentence LSTM $h_{sent}^{t-1} \in \mathbb{R}^D$, where D is the dimension of the hidden state. We calculate the weight of the channel attention β^l by the formula 1:

$$\beta^l = F_c(h_{sent}^{t-1}, V^l) \quad (1)$$

where F_c represents the channel attention function, and the calculation method of F_c is introduced in Sec. III-B.1. After that, we calculate spatial attention weight α^l through formula 2:

$$\alpha^l = F_s(h_{sent}^{t-1}, f_c(V^l, \beta^l)) \quad (2)$$

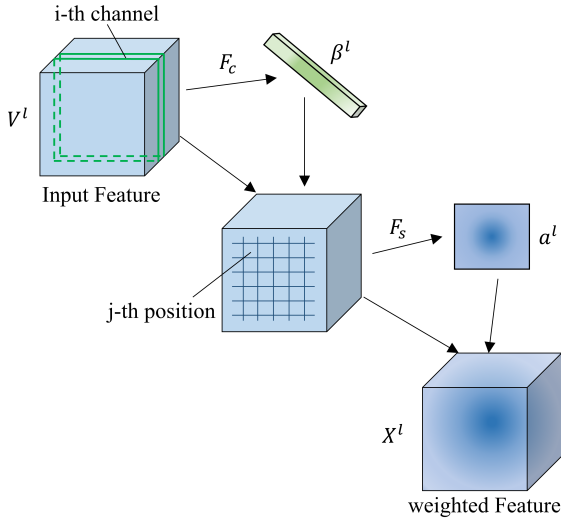


FIGURE 3. The overview of our proposed multi-attention.

where F_c represents the spatial attention function, and $F_c(\cdot)$ is the multiplication of feature map and corresponding channel weights. The calculation method of F_s is introduced in Sec. III-B.2. After we have the channel attention weights β^l and spatial attention weights α^l , we calculate the refined feature map X^l based on β^l , α^l and V^l :

$$X^l = g(X^l, \alpha^l, \beta^l) \quad (3)$$

where $g(\cdot)$ is a linear weighting function, which applies element-wise multiplication.

1) CHANNEL ATTENTION

Since each channel of the feature map is considered as a feature detector [32], it makes meaningful for the channel attention to focus on “what” for a given input image. We generate a channel attention map by exploiting the inter-channel relationship of features. Without loss of generality, we have omitted the superscript l which represents the number of layers. For the feature map V , we first reshape V to U . Then $U = [u_1, u_2, \dots, u_C]$, where $u_i \in \mathbb{R}^{H \times W}$ represents the i -th channel of feature map V , and C is the total number of channels. In order to compute the channel attention efficiently, we squeeze the spatial dimension of the input feature map by mean pooling, and then get channel feature v :

$$v = [v_1, v_2, \dots, v_C], v \in \mathbb{R}^C \quad (4)$$

where the scalar v_i is the average of the vector u_i , representing the i -th channel feature. In the t time step of sentence LSTM, the hidden state of the previous time step is h_{sent}^{t-1} . We used a single-layer neural network and a softmax function to generate the distribution of attention on the image channel. The following is the definition of channel attention function F_c :

$$b = \tanh(W_c \otimes v + b_c) \oplus W_{hc} h_{sent}^{t-1} \quad (5)$$

$$\beta = \text{Softmax}(W'_l b + b'_l) \quad (6)$$

where $W_c \in \mathbb{R}^k$, $W_{hc} \in \mathbb{R}^{k \times d}$, and $W'_l \in \mathbb{R}^k$ are weight matrices, \otimes represents the outer product of vectors. \oplus is the addition of a matrix and a vector, and the addition between a matrix and a vector is done by adding a vector to each column of the matrix. $b_c \in \mathbb{R}^k$, $b'_c \in \mathbb{R}^k$ are biases.

2) SPATIAL ATTENTION

Different from the channel attention, spatial attention focuses on “where”, which complements channel attention. We generate a spatial attention map by utilizing the inter-spatial relationship of features. Similarly, for feature map V , we flattened the width(W) and height(H) of the original V , and reshaped $V = [v_1, v_2, \dots, v_m]$, where $v_i \in \mathbb{R}^C$ and $m = W \cdot H$. Formally, v_i can be regarded as the visual feature of the i -th position on the feature map. According to the definition of channel attention function, as for the hidden state h_{sent}^{t-1} of the previous time step of sentence LSTM, the spatial attention function F_s can be defined as:

$$a = \tanh((W_s \otimes V + b_s) \oplus W_{hc} h_{sent}^{t-1}) \quad (7)$$

$$\alpha = \text{LSTM}(W_i b + b_i) \quad (8)$$

where $W_s \in \mathbb{R}^{k \times C}$, $W_{hc} \in \mathbb{R}^{k \times d}$, and $W_i \in \mathbb{R}^k$ are weight matrices, $b_s \in \mathbb{R}^k$, $b_c \in \mathbb{R}^k$ are biases.

C. LANGUAGE GENERATION

The language generation module is composed of three sub-modules, namely a sentence RNN module, a background information fusion module and a word RNN module. The sentence RNN module is responsible for generating sentence themes based on the image property vector. The background information fusion module encodes the patient’s background information and merges it with the word embedding. The word RNN module generates the most appropriate word based on the sentence topic and the merged word embedding. The three modules are described in detail in the following sections.

1) SENTENCE RNN

Sentence RNN is a single LSTM. At each time step, the sentence RNN receives the feature vector X^l from the image decoder as input, and in turn generates a hidden state sequence $h_{sent}^1, \dots, h_{sent}^t \in \mathbb{R}^H$, where H represents the dimension of the hidden state. Each hidden state h_{sent}^i has two purposes: one is to generate a topic vector q_i for word RNN input, and the other is to generate a probability p_i for determining whether the i -th sentence is the last sentence of the paragraph, where $p_i \in [0, 1]$. Formally, the sentence RNN can be written as:

$$h_t = \text{LSTM}(h_{sent}^{t-1}, X^l) \quad (9)$$

$$q_i = \tanh(W_q h_{sent}^i + b^q) \quad (10)$$

$$p_i = \sigma(W_p h_{sent}^i + b^p) \quad (11)$$

where $W_q \in \mathbb{R}^H$, $W_p \in \mathbb{R}^H$ are weight matrices, and $b_q \in \mathbb{R}^H$, $b_p \in \mathbb{R}^H$ are biases, σ is sigmoid function. The stop

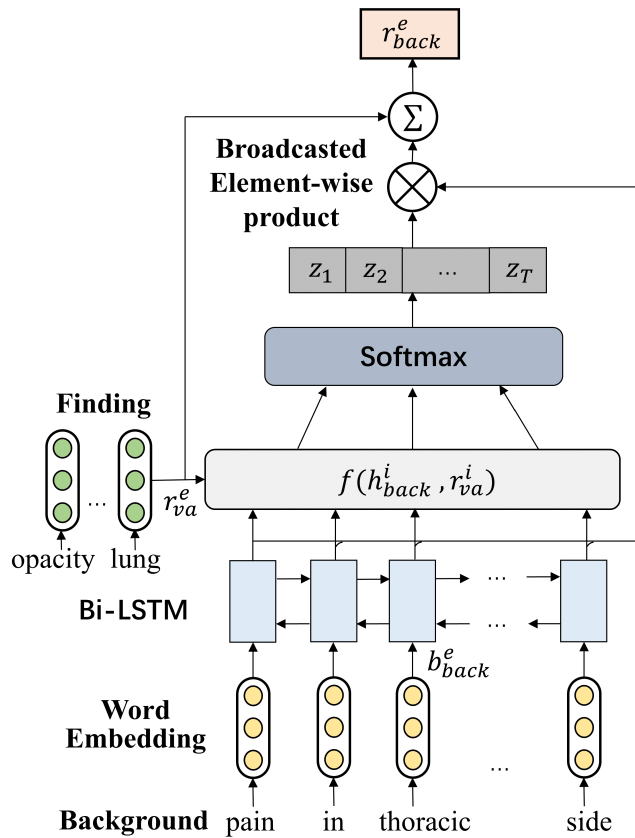


FIGURE 4. An illustration of incorporating background information definition embeddings b_{back}^e into the vanilla word embedding r_{va}^e of one report word. From the output, we will get the fused word representation r_{back}^e .

control probability p_i is greater than or equal to a predetermined threshold (eg, 0.5) indicating that the generation of the topic vector is stopped, and the word LSTM will also stop generating words.

2) INCORPORATING BACKGROUND INFORMATION DEFINITIONS

Inspired by the work of [14] and [33], we incorporated the patient's background information into the model decoder. The difference is that our approach is to fuse background information with word embedding, as shown in Fig. 4. For the "finding" and "impression" in a given report, $R = \{w_f^1, w_f^2, \dots, w_f^N\}$, where w_f^i represents a word in the "finding" and "impression", and N is the length of the text. The background information can be written as $B = \{w_b^1, w_b^2, \dots, w_b^M\}$, where w_b^i represents a word in the background information, and M is the total number of words. R and B are converted into corresponding word vectors $[r_{va}^1, \dots, r_{va}^N]$ and $[b_{back}^1, \dots, b_{back}^M]$ by pre-trained word embedding-GloVe [34]. We use a bidirectional long-term memory (BiLSTM) network to encode background information embedded in forward and backward directions. At the t -th time step, b_{back}^t is the input of BiLSTM, and the hidden

state of the BiLSTM output is b_{back}^t . This process can be written as:

$$h_{back}^1, \dots, h_{back}^M = \text{BiLSTM}(b_{back}^1, \dots, b_{back}^M) \quad (12)$$

For the word embedding generation of fusion background information, we first utilize multi-layer perceptron attention [35] mechanism to aggregate the outputs of BiLSTM and then add the aggregated vector to the pre-trained vanilla word embedding. Specifically, the multi-layer perceptron attention first computes the alignment scores of h_{back}^t and r_{va}^t through the $f(h_{back}^t, r_{va}^t)$ function:

$$f(h_{back}^t, r_{va}^t) = v^T \sigma(W^h h_{back}^t + W^{va} r_{va}^t) \quad (13)$$

where W^h and W^{va} are the weight matrices, and v is the weight vector. The softmax function then normalizes the alignment scores to form the vector z_i . By adding the output of attention and the vanilla word embedding, we obtain the new word embedding r_{back}^e after merging the background information, as shown in formula 14.

$$r_{back}^e = \sum_{t=1}^T z_t h_{back}^t + r_{va}^e \quad (14)$$

3) WORD RNN

Word RNN consists of a Bi-LSTM that generates words for the topic sentences for a given topic vector. According to Vinyals's[10] method, the first and second inputs of the Word RNN are topic vectors and a special <START> token. The difference is that the subsequent input is not the vanilla word embedding, but the word embedding of the fusion background information described in Sec. III-C.2. At each time step, the hidden state of the last LSTM layer is used to predict the distribution of words in the vocabulary, and a special <END> token is used to indicate the end of the sentence. After each word RNN generates the word of its topic sentence, these topic sentences are joined together to form the final generated diagnosis report.

D. OPTIMIZATION OBJECTIVE

The entire model is trained in an end-to-end manner. For a given training example (x, y) , x is the chest x-ray image and y is a ground-truth report description of the image, where y contains S sentences, the i -th sentences have N_i words, and $y_{(i,j)}$ represents the j -th word of the i -th sentence. After obtaining the image feature X^l , the sentence RNN unrolls S steps to simultaneously generate the topic vector and the $\{CONTINUE, STOP\}$ state distribution p_i for each sentence. The word RNN receives the topic vector and unrolls the N_i time step to generate a distribution $p_{(i,j)}$ for each word. The loss function of our model is the cross-entropy weighted sum of the stop distribution p_i of the sentence loss ℓ_{sent} and the word distribution $p_{(i,j)}$ of the word loss ℓ_{word} , defined as

TABLE 1. The number of sentences in all reports of the open-i dataset and the number of words in the sentence.

	Number of sentences				Number of words			
	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median
Finding	18	1	4.21	4	169	1	27.94	27
Impression	19	1	2.11	130	100	1	10.61	5
Total	33	2	6.32	6	230	7	38.55	34

follows:

$$\ell_{(x,y)} = \lambda_{sent} \sum_{t=1}^S \ell_{sent}(p_i, I[i = S]) + \lambda_{word} \sum_{t=1}^S \sum_{j=1}^{N_i} \ell_{word}(p_{i,j}, y_{i,j}) \quad (15)$$

IV. EXPERIMENTS

A. DATA COLLECTION

This paper adopts the Indiana University Chest X-ray Collection, which is a subset of Open-i (Open Access Biomedical Image Search Engine). The data set contains 3,955 radiology reports and 7,470 chest X-rays, including 2,314 abnormal reports, accounting for 58.51%, and each includes the indication, Findings, and Impression section. After counting all the reports, 517 reports were found missing the Findings section, and 34 were found missing the Impression section. The missing parts are marked as “Findings data is null” or “Impression data is null” respectively. According to the statistics of the reports, the median and mean values of the sentences were 6 and 6.32, respectively, and the median and mean values of the words were 34 and 38.55, respectively. More statistical information is shown in Tab. 1. Finally, all data are divided into a training set, validation set, and test set according to the proportion 80%, 10% and 10% respectively.

B. METRICS

For the time being, most of the research is based on the metrics of the general natural language generation tasks: BLEU [36], ROUGE-L [37], and CIDEr [38]. On the one hand, we continue to use common metrics because the effects of these metrics have been confirmed in a large number of natural language tasks, more importantly, the use of general metrics can better compare the research results of different researchers. On the other hand, previous studies have proposed performance metrics for this task, such as KA(keyword accuracy) [29] and CA(clinical accuracy) [31], such metrics have not been popularized, but in terms of effects, It does make up for the shortcomings of the general evaluation metrics, and inspired by this, we use MeSH(Medical Subject Headings) to evaluate the performance generated by the report. Specifically, we compute the MeSH accuracy (MA) metric as the ratio of the number of MeSH correctly generated by a model to the number of all MeSH in the groundtruth, where groundtruth was previously labeled by Demner-Fushman et al. [39].

TABLE 2. Dimension setting of different recurrent neural networks.

Module	Embedding	Hidden States
LSTM(Sentence RNN)	256	256
BiLSTM(Background)	150	150
BiLSTM(Word RNN)	256	256

C. DETAILS

1) IMAGE ENCODER

Since the resnet-152 [8] model has been proven to have good performance in medical image feature extraction and disease classification [40], we use the pre-trained resnet-152 model as the main body of the image encoder. In order to be consistent with the pre-trained resnet-152, we scaled the image down to 224 * 224 before feeding the image into the image encoder, and normalized it based on the mean and standard deviation of the images in the ImageNet training set. According to the description of the Sec.III-B, we have added a multi-attention module to the image encoder. Specifically, for the intermediate feature map V^l , the l layer uses the “res5c_branch2a” convolution layer of the resnet-152 model.

2) LANGUAGE GENERATION

This module contains three recurrent neural networks, namely the BiLSTM model of the word generation sub-module, the BiLSTM model incorporating the background information definition sub-module, and the LSTM model of the sentence topic generation sub-module. The dimension settings for the word embedding and hidden states of different recurrent neural networks are shown in Tab.2. We tokenized all the words in the “indication”, “finding” and “impression” sections of the dataset report and obtained 2,376 unique words. Considering that the size of the vocabulary is already very small, we decided to keep all the words and not delete the low-frequency words that appear only once or twice.

3) PARAMETERS

During the training process, Adam [41] optimizer was used for optimization with an initial learning rate of 1e-4. The mini-batch size was set to 24. We used variational dropout [42] for the input of BiLSTMs, which was set to 0.5. On the validation set, the threshold for stop control $T_{stop} = 0.5$. According to the statistics of Tab. 1, we set the maximum number of sentences $S_{max} = 7$, the maximum text length $N_{max} = 60$, which can cover more than 85% of the data in Openi.

TABLE 3. Evaluation of generated reports on our testing set using BLEU, CIDER, ROUGE and MA metrics. We compare our models with four baseline models. * indicates we re-implemented it according to the paper and used ResNet-152 as the image encoder.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	MA
CNN-RNN [15]	0.251	0.137	0.098	0.069	0.294	0.108	0.367
Soft ATT [16]	0.328	0.184	0.109	0.083	0.319	0.154	0.389
SCA-CNN [12]	0.347	0.216	0.112	0.087	0.328	0.159	0.411
Co-Attention [10]*	0.429	0.295	0.201	0.148	0.340	0.278	0.452
Ours(no-background)	0.368	0.234	0.144	0.113	0.323	0.209	0.460
Ours(no-multi-attention)	0.394	0.241	0.169	0.126	0.331	0.213	0.476
Ours(full)	0.476	0.340	0.238	0.169	0.347	0.297	0.498

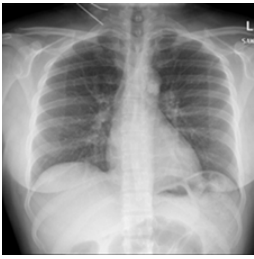


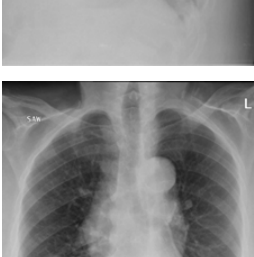
	Open report	Ours-no-background	Ours-no-multi-attention	Ours-full
	The heart and lungs have XXXX in the interval. Both lungs are clear and expanded. Heart and mediastinum normal. No change calcified aorticopulmonary XXXX node. No active disease.	There is no pneumothorax. No large pleural effusion. lungs are normal. Cardiomeastinal normal limits. No pleural effusion or pneumothorax is seen. No acute preoperative findings.	The cardiac silhouette and mediastinal contours are within normal limits. No fibrosis. No effusions. small calcified granuloma in the left lower lobe. Lungs are clear.	The heart, pulmonary are within normal limits. Heart size normal. There is no pleural effusion. Lungs are clear. There is no focal opacity. No acute cardiopulmonary disease.
	The lungs are clear. There is no pleural effusion or pneumothorax. The heart is not significantly enlarged. There are atherosclerotic changes of the aorta. Arthritic changes of the skeletal structures are noted. No acute pulmonary disease.	Cardiac and Pulmonary are normal. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. There is infiltrate with normal left lower lobe. Calcified granuloma is identified.	Silhouette process is stable in size. There is no pneumothorax. Mild prominence interstitial are unchanged. No XXXX in the bilateral lobes. No interval change. Osseous structures intact. No pneumothorax or pleural effusion.	Normal heart size and versus contours. There is no large pleural effusion. No pleural effusion or pneumothorax. Mild pectus deformity noted. No typical findings of pulmonary edema. No active disease.
	Cardiac and mediastinal silhouettes are normal. Pulmonary vasculature is normal. No pneumothorax or pleural effusion. No acute bony abnormality. XXXX XXXX opacities XXXX reflecting atelectasis versus bronchovascular crowding. Bronchovascular crowding versus atelectasis within the right lung base otherwise no acute cardiopulmonary disease.	Negative for acute displaced rib fracture. Bilateral nipple jewelry. Osseous structures unremarkable. upper mediastinal not excluded. Depending clinical suspicion mechanism investigation. Moderately narrowing spurring. Mild degenerative changes in spine. Status midline cardiomegaly.	The cardiomeastinal is midline. The lungs are clear, no pneumothorax. bony structures abnormality. The tortuous calcified aorta, stable position right atrium and ventricle. loculated pleural fluid or thickening. Degenerative changes the thoracic spine.	There are osseous abnormalities. No pleural effusion. There is a small stable XXXX foreign body noted over the left chest. Calcifications over the XXXX. cardiomegaly right lower lobe nodule and right calcifications are granulomatous process. Abnormal bilateral opacities most suggestive of atypical thickening.
	Cardiomeastinal silhouette is within normal limits for size, with redemonstration of tortuous and atherosclerotic calcified thoracic aorta. No focal consolidation, effusion, or pneumothorax identified. Eventration of the right hemidiaphragm is stable compared to prior examination. Multilevel degenerative disc disease and thoracolumbar spine again noted without acute osseous abnormality.	The heart size and pulmonary vascularity appear within normal limits. There are no nodules or masses. No effusion. Bony thorax and soft tissue is unremarkable. Negative for acute cardiopulmonary abnormality. Small nodular opacities are present in the right upper lung zone. No focal areas of pulmonary consolidation.	Heart size is normal. Enlarged calcified thoracic aorta. Scattered granulomas and bilateral perihilar calcified lymph XXXX. here are multiple reticular-nodular opacities in the upper lobes bilaterally. Increased XXXX opacities on lateral projection reflect bronchovascular crowding. Mild elevation of the right hemidiaphragm.	The thoracic aorta is calcified. There is lung calcifications from old granulomatous disease. No pleural effusion. The cardiomeastinal silhouette is within normal limits. There are multilevel degenerative changes in the spine. Increasing density in the superior of the right upper lobe. There is elevation of the right hemidiaphragm.

FIGURE 5. An example of a report generated by the Ours-no-background, Ours-no-multi-attention, and Ours-full models, where bold words represent statements that produce the same meaning as the original report.

D. BASELINES

We compared our model with several state-of-the-art image captioning methods: Vanilla CNN-RNN [15], Soft ATT [16]

and SCA-CNN [12]. We also compared it with the Co-Attention [10] model, which is currently the best way for automated radiology report generation. It is worth noting

that because the Co-ATT method did not release the original code, we re-implemented it according to the paper and used ResNet-152 as the image encoder. At the same time, in order to better reflect the effect of the Multi-Attention mechanism, we implemented a no-background model that does not incorporate the concept of background information. Similarly, we have implemented a no-Multi-Attention model that does not incorporate the Multi-Attention mechanism.

Since the open-i dataset was not released for evaluation, there is no uniform training set, validation set, and test set partitioning. Obviously, due to the small amount of data in the open-i dataset, different data partitions bring uncertainty to the results. In order to reduce the error caused by different data partitions, we performed three randomizations on the data set, so all the results in this paper are the results of averaging three different partition results.

V. RESULTS AND DISCUSSIONS

A. RESULTS

Table 1 shows the results of different models evaluated in the test set. Our full model achieved the best performance on all the evaluation metrics, which proved the effectiveness of the proposed multi-attention mechanism and the incorporating of background information into the model decoder. First of all, the use of the attention mechanism can effectively improve the performance of the model. Except for the original CNN-RNN model, all other models using the attention mechanism are better than the CNN-RNN model. The SCA-CNN model with both spatial attention and channel attention has higher evaluation metrics than the Soft ATT model with spatial attention only, and the MA metrics is 2.2% higher. Furthermore, the comparison between the Ours-full model and the Ours-no-multi-attention model using the multi-attention mechanism showed that the former's BLEU-4, CIDER and MA values were 4.3%, 9.4% and 2.2% higher, respectively.

Second, the model using a hierarchical LSTM decoder (CoAtt, no-background, no-multi-attention) is better than a model using a single LSTM decoder (CNN-RNN, Soft ATT, SCA-CNN). Among the results of Ours-no-background model and sca-cnn model, both Ours-no-background model and SCA-CNN model use Channel Attention and Spatial Attention simultaneously. The difference is that Ours-no-background using a hierarchical LSTM decoder, and BLEU-4 value increases by 3.2%. This is because the report that needs to be generated is a long text sequence containing multiple sentences, and a single LSTM will degrade performance due to gradient vanishing in tasks with too long text. Hierarchical LSTM can effectively avoid gradient vanishing by first generating sentence topics and then generating words based on topics, thereby improving the performance of the model.

Third, the comparison between the Ours-full model and the Ours-no-background model shows that the former BLEU-4, ROUGE-L, CIDER and MA values are 5.6%, 2.4%, 8.8% and 3.8% higher, respectively. This shows that the word vector

that incorporates the background information contains more representations of the patient's basic information and can effectively help the word decoder to generate words.

B. DISCUSSIONS

The Fig. 5 shows the reports generated by the three models of Ours-no-background, Ours-no-multi-attention, and Ours-full. The first line of the figure shows satisfactory results. Although there is not a lot of agreement between word order and text from the perspective of natural language, the descriptions of the three models all distinguish that the image is normal. For example: *"Both lungs are clear and expanded"*, *"lungs are normal"* and *"Lungs are clear"* all mean that there is no abnormality in the lungs, similar to *"The heart is not significant enlarged"* and *"Heart Size normal"* and so on. This can be explained. In the open-i dataset, the normal image accounts for 42%. The training of the relatively large sample makes it easier for the model to judge whether the image is normal, but at the same time different doctors have different descriptions for normal, and some descriptions are usually omitted, which makes the generated text sequence diverse.

The third line is a side chest image, the three models are very poorly effected and do not accurately describe the image information. This is related to the information carried by the lateral image itself. Usually, the side image is used to supplement the information of the missing field of the frontal image in the process of diagnosis. This result also proves that it is not feasible to use the lateral image alone to generate the report. The results of the fourth row show that Ours-full model can detect more anomalies than Ours-no-background and Ours-no-multi-attention. Compared with the Ours-no-background model, Ours-full model and Ours-no-multi-attention model can accurately point out *"elevation of the right hemidiaphragm"*, which is related to the patient's past medical history in the background information, which proves that the word embedding module incorporating background information can effectively improve the efficiency of the model. Finally, we noticed that for the generation of abnormal report, the text length we generated is usually shorter than the original report, which is related to the sentence vector "STOP" value we set according to the average number of sentences in all samples, which is why the normal report generation is better than the abnormal report generation.

VI. CONCLUSION

In this paper, we study how to automatically generate textual reports of medical images in order to reduce the workload of radiologists and help them make decisions. Our approach address two major challenges: (1) how to enhance the mapping between text and image entity position during the long text generation process of image paragraphs, (2) how to incorporate background information from different patients into the report generation model. To cope with these challenges, we propose a novel multi-attention hierarchical model that simultaneously focuses on the image's channels and spatial

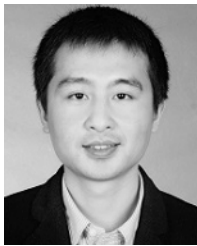
information and maps to the sentence topics. At the same time, we propose a method of word embedding that incorporates patient background information, enhancing the fusion of vanilla word embedding and background information.

Our experiment results demonstrated that the model outperforms all baselines, achieving the state-of-the-art performance in natural language metrics due to the attention and the contribution of background information fusion. We also noticed in the case analysis that, overall, the generated report is still significantly different from the original report. The reason for this gap is due to the fact that the currently published data set is too small and on the other hand because of the same Diagnosis, there are differences in the descriptions of different radiologists. Therefore, in future research, we will focus on the emergence of larger, trainable data sets and examine advanced algorithms in the natural language domain to address the problems associated with describing differentiation.

REFERENCES

- [1] A. Brady, R. 6. Laoie, P. McCarthy, and R. McDermott, "Discrepancy and error in radiology: Concepts, causes and consequences," *Ulster Med. J.*, vol. 81, no. 1, pp. 3–9, 2012.
- [2] K. C. Santosh and S. Antani, "Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?" *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1168–1177, May 2018.
- [3] S. Vajda, A. Karargyris, S. Jaeger, K. C. Santosh, S. Candemir, Z. Xue, S. Antani, and G. Thoma, "Feature selection for automatic tuberculosis screening in frontal chest radiographs," *J. Med. Syst.*, vol. 42, no. 8, p. 146, 2018.
- [4] A. Karargyris, J. Siegelman, D. Tzortzis, S. Jaeger, S. Candemir, Z. Xue, K. C. Santosh, S. Vajda, S. Antani, L. Folio, and G. R. Thoma, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays," *Int. J. Comput. Assist. Radiol. surgery*, vol. 11, no. 1, pp. 99–106, 2016.
- [5] K. C. Santosh, S. Vajda, S. Antani, and G. R. Thoma, "Edge map analysis in chest X-rays for automatic pulmonary abnormality screening," *Int. J. Comput. Assist. Radiol. surgery*, vol. 11, no. 9, pp. 1637–1646, 2016.
- [6] F. T. Zohora and K. C. Santosh, "Circular foreign object detection in chest X-ray images," in *Proc. Int. Conf. Recent Trends Image Process. Pattern Recognit.* Singapore: Springer, 2016, pp. 391–401.
- [7] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3362–3371.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, vol. 1, no. 2, pp. 2261–2269.
- [10] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2577–2586.
- [11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [12] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5659–5667.
- [13] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 6381.
- [14] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," in *Proc. EMNLP*, 2018, pp. 204–213.
- [15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.
- [16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
- [18] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1473–1482.
- [19] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 375–383.
- [20] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015, *arXiv:1511.06732*. [Online]. Available: <https://arxiv.org/abs/1511.06732>
- [21] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7008–7024.
- [22] Z. Yang, Y. Yuan, Y. Wu, R. Salakhudinov, and W. W. Cohen, "Review networks for caption generation," 2016, *arXiv:1605.07912*. [Online]. Available: <https://arxiv.org/abs/1605.07912>
- [23] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.
- [24] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 317–325.
- [25] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 747–763.
- [26] L. Melas-Kyriazi, A. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 757–761.
- [27] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.
- [28] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1530–1540.
- [29] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 457–466.
- [30] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," 2019, *arXiv:1903.10122*. [Online]. Available: <https://arxiv.org/abs/1903.10122>
- [31] 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, *arXiv:1904.02633*. [Online]. Available: <https://arxiv.org/abs/1904.02633>
- [32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.
- [33] X. Huang, Y. Fang, M. Lu, Y. Yao, and M. Li, "An annotation model on end-to-end chest radiology reports," *IEEE Access*, vol. 7, pp. 65757–65765, 2019.
- [34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
- [35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: <https://arxiv.org/abs/1409.0473>
- [36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
- [37] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, Jul. 2004, pp. 74–81.

- [38] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4566–4575.
- [39] D. Demner-Fushman, S. E. Shooshan, L. Rodriguez, S. Antani, and G. R. Thoma, "Annotation of chest radiology reports for indexing and retrieval," in *Multimodal Retrieval in the Medical Domain*. Cham, Switzerland: Springer, 2015, pp. 99–111.
- [40] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*. [Online]. Available: <https://arxiv.org/abs/1801.09927>
- [41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [42] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2575–2583.



XIN HUANG was born in 1984. He received the M.S. degree from Nanchang University, in 2010. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, China. He is currently with the Software College, Jiangxi Agricultural University, Nanchang, China. His research interests include image processing, data fusion, and machine learning.



FENGQI YAN was born in 1978. He received the M.S. degree from the Shandong University of Science and Technology, in 2007. He is currently pursuing the D.Eng. degree in electronics and information with Tongji University, China. His research interests include medical big data and medical information services.



WEI XU received the combined M.D. and Ph.D. degree from Tongji University, in 2016. She is currently an Associate Chief Doctor with the Department of Ophthalmology, Tongji Hospital, Tongji University School of Medicine, Shanghai, China. Her main research interests include optometry, big data analytics, and intelligent systems with applications to ophthalmology.



MAOZHEN LI received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, in 1997. He is currently a Professor with the Department of Electronic and Computer Engineering, Brunel University London, U.K. He has over 160 research publications in these areas including four books. His main research interests include high-performance computing, big data analytics, and intelligent systems with applications to smart grid, smart manufacturing, and smart cities. He has served over 30 IEEE conferences and is on the Editorial Board of a number of journals. He is a Fellow of the British Computer Society and the IET.

...