Historical Report Assist Medical Report Generation

Shan Ye[1], Mei Wang[1] and Yijie Dong[2]
[1]*School of Computer Science and Technology, Donghua University, Songjaing, Shanghai, China*
[2]*Department of Ultrasound, Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China*

Keywords: Automatic Report Generation, Historical Report, Encoder-decoder, Co-attention.

Abstract: How to automatically generate diagnostic reports with accurate content, standardized structure and clear semantics, brings great challenges due to the complexity of medical images and the detailed paragraph descriptions for medical images. The structure and the semantic contents of the historical report are very helpful for the current report generation. This paper proposes a text report generation method assisted by historical reports. In the proposed method, both the previous report and the keywords generated from the current images are modeled by using two encoders respectively. The co-attention mechanism is introduced to jointly learn the historical reports and the keywords. The decoder based on the co-attention is used to generate a long description of the image. The progress that learns from the historical report and the current report in the training set helps to generate an accurate report for the new image. Furthermore, the structure in the historical report helps to generate a more natural text report. We conducted experiments on the practical ultrasound data, which is provided by a prestigious hospital in China. The experimental results show that the reports generated by the proposed method are closer to the reports generated by radiologists.

1 INTRODUCTION

Medical imaging plays a crucial role in the diagnostic management and medical treatments, and imaging inspection has become a very common inspection method. Imaging doctors need to browse numerous images and write diagnostic reports with accurate content, standardized structure and clear semantics, which brings great challenges and workloads to doctors' work. In recent years, artificial intelligence especially deep learning has been making tremendous progress in various tasks. Deep learning also provides more possibilities for the automatic generation of medical reports. The automatic generation of medical reports includes two steps, namely, understanding the content of the image and generating natural language text describing the content of the image. Such a generation process is well suited to the encoder-decoder framework. In the encoder step, features of the images are extracted by using common conventional neural structures (such as AlexNet, ResNet, Inception, etc.), and then in the decoder step, the recurrent neural network generates the corresponding long text description based on the obtained features extracted by the encoder. The above process has been improved by attention mechanism (Mnih et al., 2014) or knowledge base (Li et al., 2019). However, the quality of the generated report is still unsatisfactory. There are two reasons. First, a medical report always consists of several sections describing medical observations in detail, which is a very long sequence. Take the thyroid ultrasound report as an example, about 24349 reports consist of more than 100 words. It is hard to model very long sequences and generate accurate, smoothing paragraph description by using the existing methods. More importantly, most existing works potentially learn to establish the connection between images and keywords. However, even for experienced specialists, the process of medical image interpretation can be error-prone. For example, in thyroid nodule diagnosis, calcification is an important feature. However, it is not easy to distinguish between micro-echoic focus and micro-calcification. Due to the limitation of the resolution of the instrument, the cognitive and judgment ability of the diagnostician, and many other factors, the micro-echoic foci within many nodules were misjudged as microcalcifications. It is difficult for learning models to build the correct connections between keywords and images, which leads to inaccurate reports.

In a real situation, the **dynamic change** of some features such as edge, size, calcification, echo in thyroid ultrasound can help radiologists to diagnose and accurate generation of the current report. It means

Ye, S., Wang, M. and Dong, Y.
Historical Report Assist Medical Report Generation.
DOI: 10.5220/0010245601660174
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 166-174
ISBN: 978-989-758-490-9

the historical image report is of great significance to the generation of the patient's current image report. As observed in the department of radiology, if the visual inspection is not enough to detect, describe, and classify findings in medical images, radiologists often open the patient's most recent previous report, pay special attention to the description of the abnormal and suspicious areas. The inherent disease progress helps them to obtain an accurate diagnosis and description of the current report. In fact, many diseases are chronic diseases. The patient has multiple imaging reports. Taking the thyroid nodules as an example, based on the statistics of the hospital in the past 10 years, nearly 60% of patients have more than 2 reports.

This paper proposes an automatic medical report generation method assisted by historical reports. Figure 1 illustrates the basic idea of the proposed method. Both the structure and the contents of the previous report are exploited to generate the current report. Our method adopts the encoder-decoder framework. The most recent previous report and the keywords generated from the current images are modeled by using two encoders respectively. The co-attention mechanism is introduced to jointly learn both the background information implied in the historical reports and the abnormal information implied in the keywords. The decoder based on the co-attention is used to generate a long description of the image. The progress learns from the historical report and the current report in the training set helps to generate a more accurate report for the new image. On the other hand, the text structure in the previous report has a great correlation with the current report, so a more natural text report can be generated with the help of the previous report in our method.

The main contributions of this work are:

- We propose a new medical report generation method. The historical report structure and semantics are both exploited in the proposed method.
- We introduce a structure with two encoders and one decoder. The historical report and image information are modeled by two encoders respectively. The co-attention mechanism is further provided to joint learn the historical reports and the keywords.
- We conducted the experiments based on practical ultrasound texts from the thyroid ultrasound examination data of a prestigious hospital in China to test the proposed system. The experimental results show that the proposed solution can generate more accurate and smoothing reports.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 introduces the proposed method. Section 4 presents the experimental results and Section 5 concludes the paper.

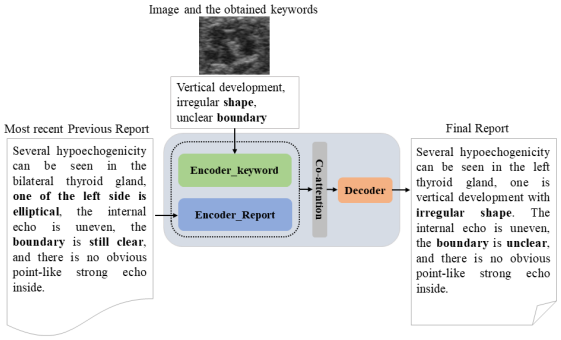


Figure 1: The basic idea of the proposed method. Both the structure and the contents of the previous report can help to generate the current report.

2 RELATED WORK

For medical report generation, one possibility is to generate the report based on the templates. In these approaches, tags and labels are learned first, then the target report is generated based on the predefined templates. The previous work proposed to segment medical images by using a semi-automatic segmentation method. Then the support vector machine classifies the image segments to get tags. Finally, the report is generated by embedding tags into candidate template sentences(Kisilev et al., 2015a; Kisilev et al., 2015b). The above methods were improved by using convolutional neural networks to obtain tags from the medical images (Kisilev et al., 2016). However, the template-based report generation method depends heavily on the image feature extraction, template sentence selection, and grammar rules. The generated report often has problems with single sentence structure, low fluency, and incoherent content.

With the development of deep learning, the encoder-decoder framework is widely used in medical report generation(Cho et al., 2014). Encoder and decoder can choose to use various convolutional neural networks and recurrent neural networks including RNN(Zaremba et al., 2014), LSTM(Hochreiter and Schmidhuber, 1997), GRU(Cho et al., 2014), BiRNN(Schuster and Paliwal, 1997) units. The encoder-decoder method was used to generate the descriptions of medical images(Shin et al., 2016), which exploit NIN architecture(Lin et al., 2013) as encoder and LSTM/GRU(Hochreiter and Schmidhuber, 1997; Cho et al., 2014) as decoder. However, the text generated by the model is the combination of multiple key-

words, not a coherent diagnostic report. The above method was improved by using the attention mechanism(Zhang et al., 2017). The attention mechanism can help the decoder module locate the most relevant encoder module output (context vector) in each step of the decoding process. A multi-task learning framework was proposed that jointly performs the prediction of medical image tags and the generation of the diagnostic report(Jing et al., 2017). And a hierarchical recurrent neural network was used as the decoder(Xue et al., 2018). The advantages of template generation methods and deep learning generation methods are combined in subsequent work(Li et al., 2018).

Attention mechanism (Mnih et al., 2014) is a commonly used method to improve the effectiveness of the deep learning models. The attention mechanism is exploited to improve the performance of report generation(Jing et al., 2017; Wang et al., 2018; Xue et al., 2018). Models that combine technologies such as reinforcement learning or knowledge graphs have also achieved excellent generation effects. The knowledge graph was incorporated into the report generation(Li et al., 2019). The visual features of the medical images were transformed into a structured abnormality graph by incorporating prior medical knowledge.

Although the above methods have improved the performance of the model in various ways, the impact of historical information is not considered when generating the report. For various chronic diseases in medicine, the historical information of patients has an important role as a reference in diagnosing the disease at the current time.

3 THE PROPOSED METHOD

3.1 Framework

The input to the proposed model is a pair of text, including the keyword set obtained from the image and the previous report of the given patient. Each keyword set consists of multiple keywords. Each keyword describes the attribute name or the attribute value which observed from the medical image. The output is the report corresponding to the given medical image $z = \{z_1, \ldots, z_n\}$, where n is the report length. Figure 2 illustrates the overall architecture of our model which consists of four modules: (a) keyword module extracts context-related and potential information from the keyword set. (b) report module learns the underlying structure and semantic information from the historical report. (c) attention module takes the hidden states of the keyword module and report module at all time steps as input to generate the weighted average context vector. (d) report output module produces the diagnostic report given the weighted average context vector.

3.2 Encoder 1: Report Module

The input to report module is historical report $t^r = \{t_1^r, t_2^r, ...t_p^r\}$, p is the length of the historical report. First, we embed historical report text input into embedding vectors. Since the historical report has a long length and carries rich semantic information. A bidirectional recurrent neural network with GRU units (Cho et al., 2014) is used as the encoder. BiRNN (Schuster and Paliwal, 1997) connects two hidden layers of opposite directions to the output. The model is therefore able to exploit information both from the past and the future. The complete context information of the historical report is hoped to be learned in this module, in which the forward hidden states is computed as follows:

$$\overrightarrow{H}^r = [\overrightarrow{h}^r_1, \overrightarrow{h}^r_2, \ldots, \overrightarrow{h}^r_p]. \quad (1)$$

We also obtained the backward hidden states:

$$\overleftarrow{H}^r = [\overleftarrow{h}^r_1, \overleftarrow{h}^r_2, \ldots, \overleftarrow{h}^r_p]. \quad (2)$$

Then, a fully connected layer is used to concatenate the forward and backward hidden states $\overrightarrow{h}^r_i$ and $\overleftarrow{h}^r_i$. To obtain the hidden state H^r_i, then we have:

$$H^r = [H^r_1, H^r_2, \ldots, H^r_p] \quad (3)$$

where H^r_i represents the bidirectional hidden layer state at i-th time step.

3.3 Encoder 2: Keyword Module

Keyword module also adopts the bidirectional recurrent neural network with GRU units (Cho et al., 2014). The input to the keyword module is the keyword set $t^k = \{t_1^k, t_2^k, ...t_l^k\}$, where l is the length of t^k. We first use an embedding layer to convert multiple keyword input into the corresponding embedding vector. The corresponding hidden layer state H^k generated by keyword module is:

$$H^k = Keyword\_encoder(embedding(t^k)) \quad (4)$$

where $H^k = [H_1^k, H_2^k, \ldots, H_l^k]$, H_i^k represents the bidirectional hidden layer state at i-th time step in keyword module.

3.4 Attention Module

By introducing the historical report information, radiologists could concentrate on observing the abnormal areas of the current medical image. The hidden

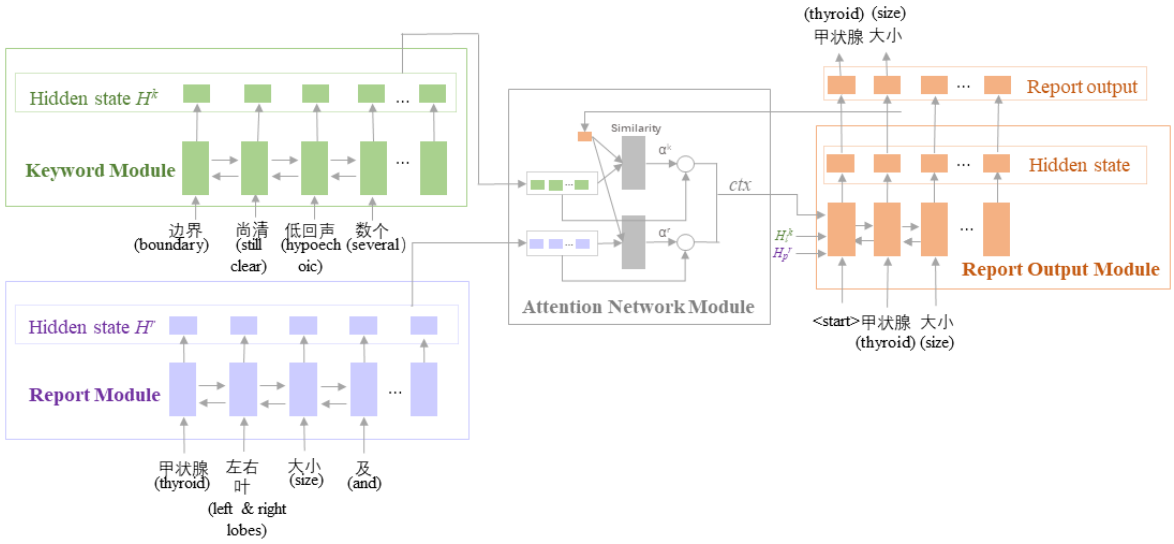


Figure 2: Overall architecture of the proposed model. Encoder 1 is the keyword module which takes multiple keyword set as input. Encoder 2 is the report module that uses the patient's historical diagnosis report as input.

state generated from the keyword module mainly provides abnormal information for the generation report. The historical reports can provide potential template information for the generation report. And it is also expected to provide background information for the generation report. For example, when there is a "border blur" in the keyword set, the "bounder blur" will be mentioned in the generation report. However, the "surface" attribute related to the border that is not mentioned in the keyword set, can be supplemented from the sentence "the boundary is clear, the surface is smooth" in the historical report.

To capture such dependencies between keyword set and historical report, we make use of the attention module as shown in Figure 2. The input to attention module is the hidden layer states at all time steps of the keyword module H^k and the report module H^r. s_{m-1} is the decoder GRU units (Cho et al., 2014) hidden state at time step m-1. At first, H^k and s_{m-1}, H^r and s_{m-1} are used to calculate the alignment scores e^k_{m-1} and e^r_{m-1} respectively:

$$e^k_{m-1} = W_{k_{att}} \cdot tanh(W_k H^k + W_{k,n} s_{m-1} + b_{ek}) + b_k,$$
$$e^r_{m-1} = W_{r_{att}} \cdot tanh(W_r H^r + W_{r,n} s_{m-1} + b_{er}) + b_r. \quad (5)$$

e^k_{m-1} and e^r_{m-1} measure how well the keywords and the historical report around position *"j"* and the output at position *"m-1"* match. For example, the higher the score e^k_{m-1}, s_{m-1} and the hidden layer states of the corresponding time steps in H^k are more similar. Then the decoder module will pay attention to the corresponding keyword observed from the image at the time step m.

Next, we apply the softmax activation function to the alignment scores to obtain the attention weights.

$$\alpha_k = \frac{exp(e^k_{m-1})}{\sum exp(e^k_{m-1})}, \quad \alpha_r = \frac{exp(e^r_{m-1})}{\sum exp(e^r_{m-1})} \quad (6)$$

where $W_{k_{att}}, W_k, W_{k,n}, b_{ek}, b_k$ are parameters of the keyword part of attention network. $W_{r_{att}}, W_t, W_{r,n}, b_{er}, b_r$ are parameters of the report part of attention network. Then, we can obtain the context vector as follows:

$$V^k = \sum_{n=1}^{N} \alpha_{k,n} \cdot H^k, \quad V^r = \sum_{l=1}^{L} \alpha_{r,n} \cdot H^r \quad (7)$$

V^k, V^r is the corresponding context vectors of keyword module and report module. The joint context vector can be obtained as follows.

$$ctx_m = [V^k_m : V^r_m] \quad (8)$$

where [:] indicates vector concatenation.

3.5 Report Output Module

Another recurrent neural network with GRU units (Cho et al., 2014) is used as the decoder in this module. Initialize the hidden layer state of the decoder with the sum of the last state of the keyword module H^k_l and the report module H^r_p. By passing the context vector ctx and all the previously predicted words $\{y_1, y_2, ..., y_{m-1}\}$ to the report output module, the decoder predicts the next word y_m:

$$p(y_m | \{y_1, y_2, \ldots, y_{m-1}\}, ctx) \quad (9)$$

At the decoding step m, the input of GRU units (Cho et al., 2014) is the joint vector of the embedding

of the previously predicted word y_{m-1} and the context vector ctx_m:

$$q_m = W_{fc} \cdot [ctx_m : y_{m-1}] \tag{10}$$

The hidden state of GRU units (Cho et al., 2014) at m-th step is calculated:

$$s_m = GRU(s_{m-1}, q_m) \tag{11}$$

Then the probability of the word y_m distribution at m-th step can be calculated as follows:

$$p(y_m|\{y_1, y_2, \ldots, y_{m-1}\}, ctx) = softmax_{y_m}(W_y s_m + b_y) \tag{12}$$

where W_{fc}, W_y, b_y are parameters and $|V_y|$ are the vocabulary size.

3.6 Training and Inference

The purpose of our model is to minimize the difference between the generation report and the report written by the radiologists. Given a training example (t^k, t^r, z), where z denotes the report written by the radiologists, our model performs encoder-decoder and produces a distribution $\hat{y}_m = p(y_m|\{y_1, y_2, \ldots, y_{m-1}\}, ctx)$ over the words. We can also obtain the ground-truth word distribution y_m by examining the presence and absence of words in z. The training loss of the model is the sparse cross-entropy losses as follows:

$$Loss = \sum Loss_i = \frac{1}{N} \sum_i - \sum_{m=1}^{v} y_{im} \cdot \log \hat{y}_{im} \tag{13}$$

where N is the size of the training set. During the training, the parameters of two encoders, decoder and attention module will be updated to the direction of lower loss through the gradient descent algorithm. The model with updated parameters has a lower loss, can generate the report which is closer to the report written by the radiologists.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Data

Our data set generated from a real-world thyroid ultrasound set from the reputable hospital in Shanghai, China, in which there are 38042 patients having thyroid ultrasound examinations. There are 21965 patients having more than one examination report. We select these patients with more than one report in our final dataset. There are 70539 ultrasound examinations reports. Also, there are 133 reports whose length exceeds 200. Since reports with long text and low proportion of reports will cause too many neurons to initialize, and most of the neurons cannot train effectively. That will increase the difficulty of the model's convergence during the training process. We remove the report with more than 200 words.

Recall that our training sample is the triple, including the most recent previous report t^r, the keyword set k^t and the current report z. So we organize each patient's report as a sequence $\{t_1, t_2, \ldots, t_m\}$ according to the report time. For the report $t_i, i \geq 2$ in the sequence, we choose t_{i-1} and t_i as t^r and z. We also extract abnormal keyword description of report t_i to obtain the keyword set t^k. In this way, we obtain 30597 triple samples in total. We divide the samples into a training set and a test set. The training set consists of 27,537 samples. There are 3060 samples in the test set.

4.1.2 Training Configuration

The dimensions of all hidden states in two encoders and one decoder are set to be 512. The dimensions in the embedding layer are set to be 256. Models are trained for 30 epochs with Adam optimizer (Kingma and Ba, 2014). The learning rate of Adam is 1e-3. The batch size is set to be 4. All models are implemented in the Tensorflow framework.

4.1.3 Baseline Methods

We compared our method with the baseline methods the neural machine translation model with attention: seq2seq_attention (Xu et al., 2015) and the related version of our model: pair2text-show-attention. The hidden state dimensions and the embedding layer dimension of the baseline models are the same as our model. In the first baseline method, we only use the keyword set to generate the report. The length of the input sequence and the output sequence in seq-to-seq are not fixed. In order to further testify the effectiveness of the proposed method, we also implement the related version pair2text_show_attention. In this version, both the historical report and the keyword set are modeled by using two encoders. While in the attention module, layers for alignment score calculation share the parameters for the two hidden state inputs.

4.1.4 Evaluation Metrics

We use the following evaluation metrics in the experiments: BLEU (Papineni et al., 2002), ChrF (Popović,

Table 1: Main results of the models on the generation report tasks. BLUE-N denotes that the BLEU score uses up to N-grams.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ChrF	NIST
seq2seq_attention	0.4815	0.4394	0.4154	0.3965	0.6086	2.3744
pair2text_show_attention	0.7629	0.7239	0.6918	0.6654	0.7892	4.8157
pair2text_two_attention	0.8084	0.7694	0.7487	0.7305	0.8621	5.1217

2015) and NIST (Doddington, 2002). BLEU is always used to measure the similarity between generated sentences and reference sentences in machine translation. Here we use BLEU measurement to measure the similarity between the report generated by the model and the report written by radiologists. N-gram overlaps are calculated in BLEU-N measurement. It has been demonstrated that BLEU is sensitive to the high-frequency words (Dugonik et al., 2014). The NIST metric is designed to improve BLEU by rewarding the translation of infrequently used words. ChrF is proposed to use character N-gram F-score for automatic evaluation, which helps to identify different word combinations.

4.2 Experimental Results

4.2.1 Quantitative Results

We compare the performance of the proposed method pair2text_two_attention with baseline methods in Table 1.

It can be seen that pair2text_two_attention significantly outperforms two baseline methods in all evaluation metrics. We can also see that pair2text_two_attention and pair2text_share_attention are both much better than seq2seq_attention method. Specifically, for BLEU-1 score, pair2text_two_attention is about 67% higher than seq2seq_attention. For BLEU-4 score, it is about 84% higher. This indicates that the previous report is very useful to generate the current report, especially it is helpful to generate longer N-grams that appear in the medical report. Also compared to seq2seq_attention, pair2text_two_attention is an increase of 41% on ChrF, which means the proposed method learns well the word combination and the context structure from the historical diagnostic report. On NIST, pair2text_two_attention has also increased by about 115% compared to seq2seq_attention, which shows that the model learned well in the keywords which appear less frequently in the training data set.

The difference between pair2text_share_attention and pair2text_two_attention is that parameters are shared in alignment score calculation. Since fewer parameters are needed to learn, pair2text_share_attention is more efficient in the training period. However, the shared parameters may ignore the difference of weight information in the historical report and keyword set. From Table 1, we can see that pair2text_two_attention outperforms pair2text_share_attention method. It is well known that the longer n-gram scores account for the fluency of the translation, or to what extent it reads ”good”. Therefore, from Bleu-1 to Bleu-4, the difficulty of evaluation gradually increases. Correspondingly the scores show a gradual downward trend. According to the table 1, the downward trend of pair2text_two_attention is the slowest. From Bleu-1 to Bleu-2, it only decreased by 4%. While for pair2text_share_attention, it decreased by 5%. It demonstrates that by using the different parameters of the attention module, the relevance of the words before and after is strongly learned.

4.2.2 Qualitative Results

Table 2 illustrates some patients examples whose reports are generated by using the proposed method pair2text_two_attention. The most previous recent report, the keyword set, the generated report and the ground-truth reported are provided in the table.

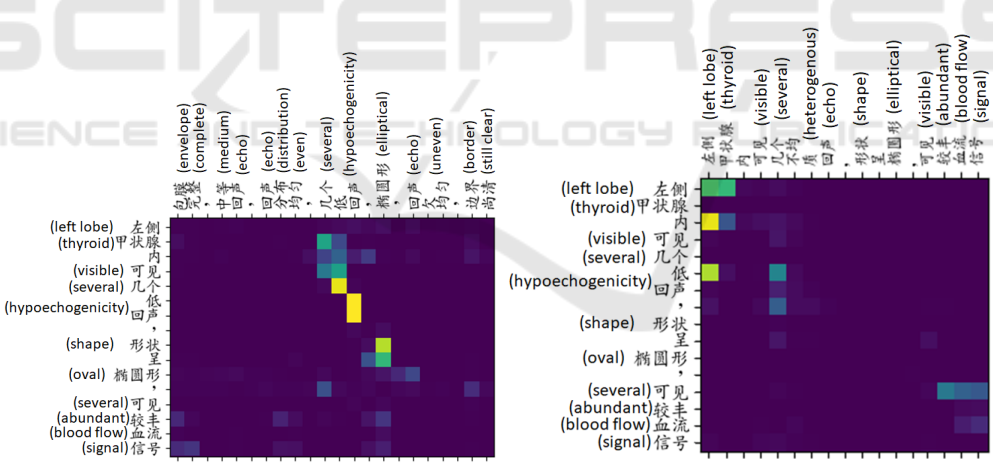
According to the first record of Table 2, our method can make good use of the information in the historical report to supply the lack of the keyword set. Comparing the historical report and the current report, it can be seen that most of the current report and the historical report have a high degree of similarity in contents and structure, only some attributes may be different. Take the first record as an example, information such as “thyroid shape, thyroid size” and “no obvious abnormality” can be learned from the historical report, and the information of “echo distribution” and “echo type” can be obtained from the keyword set. Using our model, we can use the information from the historical report and the keyword set to generate the report with clear semantics and complete structure.

Comparing the second and third records in Table 2, the keywords of the two records are the same, however since their historical reports are different, the generated reports are different and each generated report is very similar to the ground-truth report.

In the fourth record of Table 2, the pair2text_two_attention model not only accurately obtains the information from the keyword set about

Table 2: Four patient examples whose reports are generated by pair2txt_two_attention based on the input of keyword set and historical report. The first example illustrates the association between the current report with the keyword set, historical report and generated report. The second and third examples have the same keyword set, but different historical reports. The fourth example shows the trend of nodules over time.

Historical report	Keyword set	Generated report	Ground-truth report
The left and right lobes of the thyroid gland are normal in shape, the thickness of the isthmus is normal, the border is clear, the surface is smooth, the envelope is complete, and the internal echo is a dense medium echo, and the echo distribution is heterogenous. CDFI: No obvious abnormal blood flow signal. **Several hypoechogenicity can be seen in the left thyroid gland, the shape is elliptical, the internal echo is heterogenous, the border is still clear, and there is no obvious point-like strong echo inside. CDFI: an abundant blood flow signal is visible.**	Envelope complete, medium echo, echo distribution even, **several hypoechogenicity**, elliptical, echo uneven	The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, the echoes distribution is even. CDFI: No obvious abnormal blood flow signal. **Several hypoechogenicity in the left thyroid gland, the shape is elliptical, the internal echo is uneven, the boundary is still clear, and there is no obvious point-like strong echo inside. CDFI: an abundant blood flow signal is visible.**	The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, the echo distribution is even. CDFI: No obvious abnormal blood flow signal. **Several hypoechogenicity in the left thyroid gland, the shape is elliptical, the internal echo is uneven, the boundary is still clear, and there is no obvious point-like strong echo inside. CDFI: an abundant blood flow signal is visible.** ....
Postoperation, right lobe: anteroposterior diameter, left lobe (resected), isthmus: resected. In the residual thyroid, the internal echo is a dense medium echo, and echo distribution is even.	**Envelope complete, medium echo, echo distribution even**	**Postoperation, left lobe (resected), the residual right thyroid gland** is normal in size and shape, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even.	**In the residual right lobe of the thyroid gland** is normal in size and shape, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and echo distribution is even. ....
The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. ...	**Envelope complete, medium echo, echo distribution even**	**The left and right lobes of the thyroid gland** are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. ...	**The left and right lobes of the thyroid gland** are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. ...
The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. CDFI: No obvious abnormal blood flow signal. **A mixed echo in the left thyroid gland. The shape is elliptical, the internal echo is uneven, the boundary is still clear**, and there is no obvious point-like strong echo inside.	Envelope complete, medium echo, echo distribution even, **a liquid mixed echo, elliptical, echo distribution uneven point-like strong echo**	The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. CDFI: No obvious abnormal blood flow signal. **A liquid mainly mixed echo in the left thyroid gland. The shape is elliptical, the internal echo is uneven, the boundary is still clear, and there is a spot-like strong echo (colloid agglutination)**.	The left and right lobes of the thyroid gland are normal in size and shape, the thickness of the isthmus is normal, the boundary is clear, the surface is smooth, the envelope is complete, and the inside echo is a dense medium echo, and the echo distribution is even. CDFI: No obvious abnormal blood flow signal. **A liquid mainly mixed echo in the left thyroid. The shape is elliptical, the internal echo is uneven, the boundary is still clear, and there is a spot-like strong echo (colloid agglutination)**. ....



(a) Matrix_Keywords-record1 (b) Matrix_Report-record1

Figure 3: The attention weight matrix for the first example in Table 2. The row lists the generated report, the column of Figure (a) is the keyword set, and the column of Figure (b) is the historical report.

the nodule, such as "a liquid mixed echo", but also learn information from "a mixed echo left thyroid gland" in the historical report. Then the model can infer that this is a description of the same nodule at different times. The echo type of the nodule changes over time, from "mixed echo" to "liquid mainly mixed echo". In the pair2text_two_attention model, the progress of the disease is learned from the historical report and the current report in the training set.

4.2.3 Attention Weight Matrix

To further understand how the historical report and keyword set to help generate the current report, we provide the attention weight matrix when the patient examples' reports in Table 2 are generated. Since there are two attention layers in the proposed method,

each example has two attention weight matrices (denote as Matrix_Report and Matrix_Keywords in the following part) corresponding to historical report and keyword sets. By visualizing the attention weight matrix, we focus on analyzing what the proposed method learned.

Figure 3 illustrates parts of Matrix_Report and Matrix_Keywords of the first report in Table 2. In both Figure 3a and 3b, each row represents a set of weights to construct the new vector. For example, there is a higher weight between row 7 with "hypoechogenicity" and "hypoechogenicity" of keyword set in Figure 3a. Since "left lobe" does not appear in the keyword set, in the first row of Figure 3a, the weights are all close to 0. In contrast, in the first row of Figure 3b, the corresponding words "left" and " thyroid" have higher weight.

By comparing Figure 3a and 3b, it can be seen that the model learns "left thyroid" from historical reports and "several hypoechogenicity visible" from keyword set. We can also infer that potential relevance between phrases is learned, such as "left" and "thyroid". Such relevance could be exploited to estimate the probability distribution of the words to be generated. Let's see the rows in Figure 3b. the generated words "visible, abundant" and " blood, flow, signal" in the historical report have higher weights. However, the generated word "visible" and "visible" in the historical report the historical report has small weight. At the same time, the subsequent words "blood flow signal" in the generated report depends on the generated word "visible". The potential reason might be that the learned relevance influences the probability distribution calculated based on the part of the generated report.

5 CONCLUSION

This paper proposed the method that generates the current medical report both from the most recent previous report and the keyword set observed from the current medical image. The experimental results demonstrated the effectiveness of the proposed method. In the future, we plan to design a more efficient learning strategy for model inference. Also, the method that the previous report helps the generation of the keyword set is to be investigated.

ACKNOWLEDGEMENTS

This work were supported by the National Key R&D Program of China (2019YFE0190500) and the Shanghai Innovation Action Project of Science and Technology (18511102703).

REFERENCES

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Dugonik, J., Boskovic, B., Maucec, M. S., and Brest, J. (2014). The usage of differential evolution in a statistical machine translation. In *2014 IEEE Symposium on Differential Evolution (SDE)*, pages 1–8.

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

Jing, B., Xie, P., and Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv:1711.08195*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kisilev, P., Sason, E., Barkan, E., and Hashoul, S. (2016). Medical image captioning: Learning to describe medical image findings using multi-task-loss cnn. *Deep Learning for Precision Medicine*.

Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., and Hashoul, S. Y. (2015a). From medical image to automatic medical report generation. *IBM Journal of Research and Development*, 59(2/3):2–1.

Kisilev, P., Walach, E., Hashoul, S. Y., Barkan, E., Ophir, B., and Alpert, S. (2015b). Semantic description of medical image findings: structured learning approach. In *BMVC*, pages 171–1.

Li, C. Y., Liang, X., Hu, Z., and Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, volume 33, pages 6666–6673.

Li, Y., Liang, X., Hu, Z., and Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. In *NIPS*, pages 1530–1540.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth*

Workshop on Statistical Machine Translation, pages 392–395.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Shin, H., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016). Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, pages 2497–2506.

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, pages 9049–9058.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.

Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., and Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, pages 6428–6436.