Medical Report Generation with Multi-Attention for Abnormal Keyword Description and History Report

Mei Wang
wangmei@dhu.edu.cn
Donghua University
Songjiang District, Shanghai, China

Haihan Yao
947727923@qq.com
Donghua University
Songjiang District, Shanghai, China

Yanxia Qin*
yxqin@dhu.edu.cn
Donghua University
Songjiang District, Shanghai, China

ABSTRACT

This paper proposes an automatic medical report generation framework based on both current medical image and a previous history report. A keyword list describing the abnormal or special observations from the medical image is used to represent the image. In the proposed method, sentence-level structure information of the history report is extracted with the sentence level embedding. Then we construct two attention components. One is used to learn important semantic and sequential information from the keyword list, the other is used to learn the correlation between the current keyword list and the history report. Finally, all above information is combined together to help generating the current report. We conduct experiments on a practical ultrasound text dataset collected from a reputable hospital in Shanghai, China. The experimental results show that the reports generated by the proposed method are more accurate and smooth compared with a strong baseline method.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**.

KEYWORDS

medical report generation, multi-attention , keyword, history report

1 INTRODUCTION

Medical imaging plays a crucial role in the diagnostic management and medical treatments. 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 semantic meaning, which brings great challenges and workloads to doctors. In recent years, artificial intelligence especially deep learning has been making tremendous progress in various tasks [9]. Deep learning also provides more possibilities for the automatic generation of medical reports[5].

Intuitively, automatic generation of medical reports includes two steps, which are understanding the content of the medical image and generating natural language text to describe its content[12]. The process is well suited to the encoder-decoder framework[8]. In the encoder step, important features of the image are extracted using conventional neural models such as AlexNet[7], ResNet, and Inception, etc. In the decoder step, the recurrent neural network generates the corresponding long text descriptions based on features extracted by the encoder[13]. However, the quality of the generated report is unsatisfactory. There are two reasons. Firstly, accurately understanding medical images is a challenging task. Even for experienced specialists, the process of medical image interpretation can be error-prone. More importantly, medical report always consists of several sections describing medical observations in detail, which is long and has flexible structures[6]. It is hard to model very long sequences and generate accurate, smoothing paragraph to describe images by using existing methods even the keywords are generated correctly.

In real situation, the history report of a given patient has great help in generating his/her current report. As observed in the department of radiology, radiologists often use the patient's most recent previous report for reference. The knowledge implied in the disease development sequence may help doctors to confirm or correct abnormal and suspicious observations and descriptions. Besides, the text structure in the previous report provides guidance in generating the current report. In our previous work [14], we initially proposed that using two attentions to learn from the keyword list obtained by current image and the history report for current report generation. However, these two attentions are separately learned, regardless of the influence between the words in history report and the words in keyword list. In addition, the relationship between the current keyword list and the history report is ignored. In this paper, we further study how to emphasize more important words in the keyword list and also exploit the valuable relationship between the current keyword list and the history report.

Table 1 illustrates an example of history report, keyword list and current report. From the table, we can observe that three kinds of information are very important for current report generation, namely, abnormal observations in the keyword list, textural structures of the history report and the combined information in the history report and current observations. We propose to model semantics and textual structures of the history report with a BERT-based sentence encoding model. We utilize one attention-based module to emphasize abnormal keywords in the keyword list. In addition, the combined information between the keyword list and the history report is captured with another attention-based module. All these valuable information are incorporated into an encoder-decoder based framework for generating current report. Experiments on a practical ultrasound text dataset from a prestigious hospital in China verify the effectiveness of the proposed model with more accurate and smoothing generated reports. The $BLEU\_1$ and $BLEU\_4$ are achieved 83.8% and 76.7% compared with the ground truth report which are generated by clinical experts.

The rest of the paper is organized as follows. Section 2 introduces the novel attention network. Section 3 presents the experimental results and Section 4 concludes the paper.

*corresponding author

Table 1: A triple example of a history report, a keyword set and a ground-truth report. Important examination targets are indicated in red color. Attributes of organs are denoted in blue color.

History report	Keyword set	Ground-truth report
The shape of thyroid gland on both sides is normal, the capsule is smooth, the internal echo is even, and there are many large and solid nodules with different sizes in the essence. CFI shows that the blood supply of thyroid parenchyma is normal. No abnormal lymph nodes were found in both side of the neck. No abnormal mass echo was found in bilateral parathyroid region.	Irregular shape,uneven, multiple cystic solid, increased blood supply	The shape of bilateral thyroid gland was irregular, the capsule was smooth, the internal echo was diffusely reduced and uneven. Multiple cystic and solid nodules of different sizes could be seen in the parenchyma. CFI showed that the blood supply of thyroid parenchyma increased. No obvious abnormal lymph nodes were found in the both side of the neck. No obvious abnormal mass echo was found in bilateral parathyroid area.

2 THE PROPOSED METHOD

2.1 Data

We denote the data sample as a triple $D = \{r, k, o\}$, o denotes the current report, r denotes its most recent history report, k is the keyword sequence obtained. D_r, D_k, D_y denote the history report set, keyword sequence set and current report set respectively. In our dataset, there are 38042 patients having thyroid ultrasound examinations, in which 21965 patients having more than one thyroid ultrasound examination reports. For the patients with more than one reports, we organize his/her reports as a sequence $\{d_{t_1}, d_{t_2}, \ldots, d_{t_m}\}$ according to the report time t_i. For the report $d_{t_i}, i \geq 2$ in the sequence, we choose $d_{t_{i-1}}$ and d_{t_i} as h and r. In this way, we obtain 30597 triple samples in total.

2.2 Framework

Figure 1 shows the architecture of our model. The model consists of encoder layer, attention layer and the decoder layer. The encoder layer has two modules. In the encoder module, GRU [2] is used to encode the keyword sequence k into a word vector. To extract the sentence level structure, the report module uses BERT to encode each sentence from the history report into a sentence vector, which is stacked into a sentence vector information matrix. The report output module in decoder layer is a GRU structure, which is responsible for predicting words in sequence according to the output of the attention layer. It is easy to see that the semantics of the current report consists of two parts. One part is the default or regular description, the other is the abnormal or special observation. The formal information should be mainly obtained from the history information. The keyword sequence should contribute more to the second part of the report. Thus attention components are designed in the attention layer. The detailed information is introduced in the following section.

2.3 The Attention Layer

We utilize one attention-based module to emphasize abnormal keywords in the keyword list.

Assume the hidden state of the keyword module for a given keyword sequence k as $H^k = (h_1^k, h_2^k, h_3^k, ..., h_K^k)$. The output hidden state of the report module for r is denoted as $H^r = (h_1^r, h_2^r, h_3^r, ..., h_R^r)$. The alignment scores of h_j^k and the hidden state h_t^o of the report output module at time t is calculated by the typical Bahdanau attention [1]:

$$score_{context}(h_t^o, h_j^k) = V_{k_2} tanh(W_{k_3} h_t^o + W_{k_4} h_j^k) \tag{1}$$

$$\alpha_{tj}{}^{k_{context}} = \frac{exp(score_{context}(h_t^o, h_j^k))}{\sum_{j=1}^{K} exp(score_{context}(h_t^o, h_j^k))} \tag{2}$$

where α is the weight vector. As we observed, the weight distribution obtained from Eqn.2 is a little bit flat. The difference between scores are reduced. So the importance of the corresponding keywords are reduced too.

To increase the salience of the keywords, we propose a new method to combine the advantage of Content base attention[4] and local-p attention[10]. The local-p attention intercepts a window with P_t as the center and a length of $2D + 1$ on [1,K], which limits the scope of attention.It uses Gaussian function to act on the score at the end to get attention weights. The attention weights obtained in this way not only rely too much on the accurate selection of P_t, but also have nothing to do when the number of points to be emphasized is greater than or equal to 2. It implies that the data information outside the window will be completely ignored. We propose highlight attention removes this limited window, and uses an exponential function instead of the Gaussian function to achieve the effect of emphasis and weakening.

$$score_{salient}(h_t^o, h_j^k) = cosine(h_t^o, h_j^k) \tag{3}$$

$Score_{salient}$ measures the degree of similarity between h_t^o and h_j^k. The more similar they are, the closer the value is to 1. The greater the difference between them, the closer the value is to -1.

$$\alpha_{tj}{}^{k_{salient}} = \frac{exp(n \cdot score_{salient}(h_t^o, h_j^k))}{m} \tag{4}$$

By controlling the size of m and n, different exponential functions can be selected for calculation. So as to achieve different degrees of emphasis. Then, we calculate the environment vector:

$$C_t^{k_{context}} = \sum_{j=1}^{K} \alpha_{tj}^{k_{context}} \cdot h_j^k, \quad C_t^{k_{salient}} = \sum_{j=1}^{K} \alpha_{tj}^{k_{salient}} \cdot h_j^k \tag{5}$$

Then a soft attention is used to learn to balance the context and salient information. We use the vector h_t^s and C_t^k to calculate the

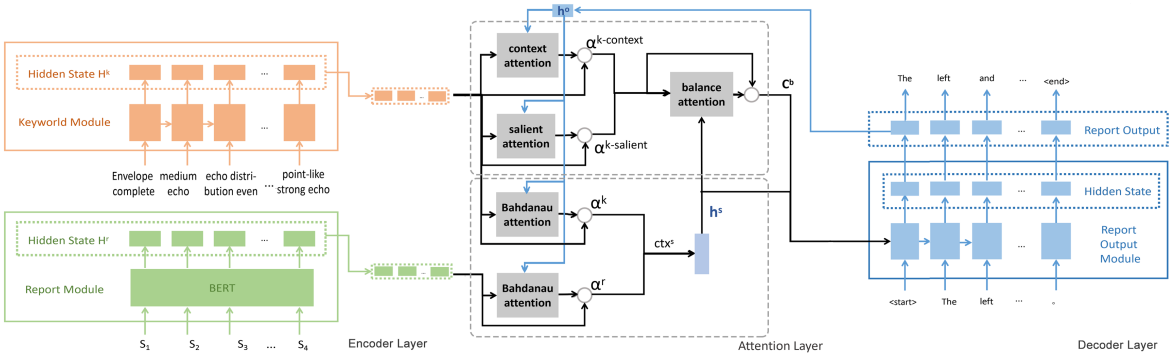

Figure 1: Overall architecture of the proposed model.

weight.

$$score(h_t^s, C_t^{k_{context}}) = V_{b_1} tanh(W_{b_1} h_t^s + W_{b_2} C_t^{k_{context}}) \quad (6)$$

$$score(h_t^s, C_t^{k_{salient}}) = V_{b_2} tanh(W_{b_3} h_t^s + W_{b_4} C_t^{k_{salient}}) \quad (7)$$

Among them, $W_{b_1}, W_{b_2}, W_{b_3}, W_{b_4}, V_{b_1}, V_{b_2}$ are all parameters.

$$\alpha_t{}^{k_{context}} = \frac{exp(score(h_t^s, C_t^{k_{context}}))}{exp(score(h_t^s, C_t^{k_{context}})) + exp(score(h_t^s, C_t^{k_{salient}}))} \quad (8)$$

$$\alpha_t{}^{k_{salient}} = \frac{exp(score(h_t^s, C_t^{k_{salient}}))}{exp(score(h_t^s, C_t^{k_{context}})) + exp(score(h_t^s, C_t^{k_{salient}}))} \quad (9)$$

The final context vector C_t^b at time t is calculated as follows:

$$C_t^b = \alpha_t{}^{k_{context}} \cdot C_t^{k_{context}} + \alpha_t{}^{k_{salient}} \cdot C_t^{k_{salient}} \quad (10)$$

The history report is composed of multiple sentences organized by certain semantic structure. Each sentence has its theme and meaning. For example, in general, the sentences which describe the nodule followed the sentences to describe the thyroid gland background. Obviously, there is a clear structural relationship between the sentences. Therefore, the sentence vector matrix from the encoder is used as the input of the second attention module. Two Bahdanau attentions are used to get relation information among keywords and sentences of history report respectively and we denote them with C_t^k and C_t^r.

A GRU network is used to learn the combined information between C_t^r and C_t^k:

$$ctx_t^s = [C_t^r : C_t^k] \quad (11)$$

$$h_t^s = GRU(ctx_t^s, h_{t-1}^s) \quad (12)$$

where h_t^s represents the output at time t.

2.4 Training and Inference

Given a training example $D = \{k, r, o\}$, 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.

The training loss of the model is the sparse cross-entropy losses as follows:

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

where N is the size of the training set.

3 EXPERIMENTS

3.1 Experimental Setup

3.1.1 Training Configuration. As mentioned above, the constructed dataset includes 38042 samples in total. We randomly choose 3000 samples for experiments, in which 80% of them are taken as training data and the other 20% as test data. For keyword list, we did not generate it directly from image observation. We extract the abnormal and special description from the current report.

The model is implemented under the Tensorflow framework. The number of hidden units in all GRU networks is set to be 256. The dimension of word embedding is set to be 256. Batch size is set to be 64. Models are trained for 80 epochs with the Adam optimizer[11]. Parameters m and n in Eq. 4 are empirically set to be 20 and 3. Sentence representations of history reports are learned with a BRET model called "roberta-base-word-chinese-cluecorpussmall"[15].

3.1.2 Baseline Methods. Two groups of experiments were conducted. In the first group, we compare the proposed model with our previous work Co_attention [14], which only consider the word level information. Here we denote the proposed model with two attention modules as $M(Sen\_att, Keyword\_att)$ In the second group, we testify the effective of each attention module. At first, the results of the model without the attention for sentence-level embedding denoted as $M(Keyword\_att)$ were provided. Then, we replace the proposed keyword attention mechanism with single Bahdanau Attention $M(Sen\_att, B\_att)$, Content Base Attention $M(Sen\_att, C\_att)$, Highlight Attention $M(Sen\_att, H\_att)$ respectively to testify the effectiveness of the proposed keyword attention mechanism.

3.1.3 Evaluation Metrics. We use BLEU[14], ChrF[14] and NIST[3] to measure the similarity between the generated diagnostic report and the target diagnostic report. As we know, the keyword sequence always reflects abnormal observations and plays a important role to represent semantics of current reports. So we further define the percentage of appeared keywords in target report to measure the accuracy of the generated report. Let the key word sequence $r = \{r_1, r_2, \ldots, r_s\}$, and s is the number of keywords in the sequence. Define function $p(r_i)$, if r_i appears in the generated

Table 2: Results of different architectures on the generation report tasks.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ChrF	NIST	$Percent_{r,o}$
Our Model $M(Sen\_att, Keyword\_att)$	0.8121	0.7758	0.7466	0.7214	0.8542	4.9537	83.36%
Co_attention Baseline	0.8001	0.7551	0.7190	0.6872	0.8330	4.8908	77.39%
$M(Keyword\_att)$	0.2655	0.1440	0.0857	0.0531	0.3451	1.3504	25.21%
$M(Sen\_att, B\_att)$	0.8150	0.7804	0.7524	0.7279	0.8595	4.9668	82.31%
$M(Sen\_att, C\_att)$	0.8081	0.7715	0.7426	0.7175	0.8593	4.9343	82.35%
$M(Sen\_att, H\_att)$	0.7964	0.7606	0.7320	0.7076	0.8521	4.8393	82.89%

report o, then $p(r_i) = 1$. The evaluation metric is calculated as $Percent_{r,o} = \sum_r \frac{\sum_{i=1}^{s} \frac{p(ri)}{s}}{N}$.

3.2 Experimental Results

Table 2 illustrates the experimental results of different methods. It can be seen that the proposed model outperforms the baseline model Co-attention in all evaluation metrics. Specifically, for BLEU-4 score, it is about 5% higher. This demonstrates that the report generated by the proposed model is more closely related to the reference report in phrase level measure. More importantly, the percentage of keywords appear in the report increased from 77.39% to 83.36%. Since keywords are the crucial information in the reference report and important prompts for generating the report, the presence or absence of keywords in the generated report reflects a considerable part of the correctness of the generated report. Under ideal circumstances, keywords should appear in the generated report 100%. The significant increase of this evaluation indicator means that the semantic accuracy of the generated report is improved.

Several other experiment results show the effectiveness of different network structures and different attention. It is easy to see that $M(Keyword\_att)$ has poor-performance. It means that the attention to learn sentence level embedding plays a very important role in the network. In the next three row of Table 2, we can see that the *BLEU* is declined, which means the sentence similarity between the generated report and the reference report decreases, while the percentage of keywords' appearance in the report increases. This result tells us that highlight attention has advantages in obtaining the salient information of keyword sequence, which can make more keywords appear in the generated report and increase the semantic accuracy of the report. Bahdanau attention is relatively better at extracting context information, which can make the generated report closer to the terminology of the reference report. It can also be seen that the BLEU, ChrF, and NIST of the proposed Model are closer to the Bahdanau attention model. While the percentage of keywords that appear in the the report is highest. This proves the effectiveness of the proposed method.

4 CONCLUSION

This paper proposed a multi-attention model which helps to incorporate the sentence level structure and the word level correlation learned from the history report to generate the current report. The experiments conducted on the real world dataset show the effectiveness of the proposed method. In the future, we plan to apply our method to different medical image datasets and try to represent the structure of history reports with more accurate model such as tree-based model.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant (No.2019YFE0190500), the Shanghai Innovation Action Project of Science and Technology (No.18511102703) and the NSFC project (No. 62006039).

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:abs/1409.0473
[2] K. Cho, B. Van Merrienboer, 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
[3] 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*. 138–145.
[4] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing Machines. arXiv:1410.5401
[5] P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul. 2015. From medical image to automatic medical report generation. *IBM Journal of Research and Development* 59, 2/3 (2015), 2–1.
[6] P. Kisilev, E. Walach, S. Y. Hashoul, E. Barkan, B. Ophir, and S. Alpert. 2015. Semantic description of medical image findings: structured learning approach.. In *BMVC*. 171–1.
[7] Hinton G E. Krizhevsky A, Sutskever I. 2012. ImageNet classification with deep convolutional neural networks.. In *International Conference on Neural Information Processing Systems*. Curran Associates Inc., 1097–1105.
[8] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6666–6673.
[9] Y. Li, X. Liang, Z. Hu, and E. P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in neural information processing systems*. 1530–1540.
[10] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention based neural machine translation. arXiv:1508.04025
[11] Kingma D. P. and Ba J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980
[12] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9049–9058.
[13] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 457–466.
[14] Shan Ye, Mei Wang, and Yijie Dong. 2021. Historical Report Assist Medical Report Generation. In *HEALTHINF 2021: International Conference on Health Informatics*. Vienna, Austria.
[15] Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An Open-Source Toolkit for Pre-training Models. arXiv:1909.05658