Mathematical Biosciences and Engineering

MBE, 19(11): 11768–11781.
DOI: 10.3934/mbe.2022548
Received: 04 July 2022
Revised: 13 July 2022
Accepted: 20 July 2022
Published: 16 August 2022

http://www.aimspress.com/journal/MBE

Research article

Research on chest radiography recognition model based on deep learning

Hui Li[1], Xintang Liu[1], Dongbao Jia[1,*], Yanyan Chen[1], Pengfei Hou[1] and Haining Li[2,*]

[1] School of Computer Engineering, Jiangsu Ocean University, China
[2] Department of Neurology, General Hospital of Ningxia Medical University, China

* **Correspondence:** Email: dbjia@jou.edu.cn, 15340368@qq.com.

Abstract: With the development of medical informatization and against the background of the spread of global epidemic, the demand for automated chest X-ray detection by medical personnel and patients continues to increase. Although the rapid development of deep learning technology has made it possible to automatically generate a single conclusive sentence, the results produced by existing methods are not reliable enough due to the complexity of medical images. To solve this problem, this paper proposes an improved RCLN (Recurrent Learning Network) model as a solution. The model can generate high-level conclusive impressions and detailed descriptive findings sentence-by-sentence and realize the imitation of the doctor's standard tone by combining a convolutional neural network (CNN) with a long short-term memory (LSTM) network through a recurrent structure, and adding a multi-head attention mechanism. The proposed algorithm has been experimentally verified on publicly available chest X-ray images from the Open-i image set. The results show that it can effectively solve the problem of automatic generation of colloquial medical reports.

Keywords: medical; recurrent learning network (RCLN); convolutional neural network (CNN); long short-term memory (LSTM)

1. Introduction

The reading and interpretation of medical images is usually performed by medical professionals. But even for experienced experts, this process of medical image interpretation and reporting is also prone to error. Staff shortages and overworked work-loads can also lead to misjudgments in radiology reports. Writing accurate medical imaging reports is necessary for inexperienced radiologists and pathologists, especially in rural areas and in areas where the quality of care is relatively low. For

experienced radiologists and pathologists, writing imaging reports is tedious and time-consuming, and automated generation of medical reports can effectively reduce the doctors' workload and mistakes.

There are several issues that must be addressed to automate the generation of auxiliary reports. First, a complete diagnostic report consists of several different forms of information. Second, how to locate the image region and describe it correctly. You et al. [1] automatically extracted machine-learnable annotations from regression data, but the description results were still not ideal. Third, the description in the image re-port contains multiple sentences. Krause et al. [2] used the combined structure of image and language to generate hierarchical descriptive paragraphs, while generating such a long text is still relatively difficult to achieve. Fourth, the automatically generated statements are still unreadable and cannot be colloquial in a human voice. The current single-layer LSTM method cannot model long word sequences. The traditional RNN-+CNN architecture is difficult to generate long statement sequences. Multimodal recurrent model with attention (MRNA) can be used to model long word sequences, but the accuracy is very low and lacks readability.

In view of the above problems, the following conclusions are drawn. 1) The verbal information of medical reports is more important than the image information. 2) The final results are often more concerned with the degree of imitation of the doctor's tone. Based on this, RCLN model is proposed in this paper. RCLN model solves the problem of multiple forms of information by establishing a multi-task framework. On the area localization problem, a research team proposed a new real-time automatic calibration scheme based on scanning sources. The proposed method allows accurate calibration regardless of the path length variation caused by the non-planar topography of the sample or the scanning of the galvanometer [3]. Previously, the application of multimodal imaging technology in the study of density changes of melanosomes and lipofuscin granules in the retinal pigment epithelium (RPE) cells [4]. There is also an efficient direct time-domain resampling scheme based on phase analysis, which shows significant performance improvements in terms of accuracy and speed and silica-coated silver nanostructures can be excellent contrast agents for optical coherence tomography (OCT) imaging [5]. Multi-label classification is a multi-label classification task processing model, it regards label prediction as a multi-label classification task and long description generation as a text generation task. To solve the problem of image region localization, the MRNA model introduced a cooperative attention mechanism, and explored the synergistic effect of visual features and semantics in the grouping while biased towards im-ages and prediction labels. In view of the difficulty in generating long text, RCLN uses hierarchical LSTM to induce long text by taking advantage of the constituent nature of reports. Combined with the cooperative attention mechanism, the hierarchical LSTM first generates high-level topics, and then generates fine-grained descriptions according to the topics.

1) Aiming at the confusion of long sentences in traditional medical report generation and the difficulty in locating diseased areas, a new cycle sentence generation model and LSTM word-by-word generation model with attention were proposed to solve the problems of long text and colloquialism and achieve theoretical innovation.

2) Through comparative experiments, it is proved that the model is more effective than the traditional model in the generation of chest X-ray reports.

2. Related technologies

2.1. Problem definition

First of all, the first task is to predict the label of a given image. The label prediction task is processed in the way of multi-label classification task. Specifically, features of the given image I are firstly extracted:

$$\boldsymbol{p}_{\boldsymbol{1},pred}\left(\boldsymbol{l}_i = 1 \mid \{\boldsymbol{v}_n\}_{n=1}^{N}\right) \propto \quad exp\left(\boldsymbol{MLC}_i(\{\boldsymbol{v}_n\}_{n=1}^{N})\right) \tag{1}$$

where I ∈R, L is the label vector, li = 1/0 indicates whether there is the i_{th} label, and MLC_i represents the i_{th} output of the network. A complete diagnostic report is composed of multiple internal reports with different forms of information. The chest X-ray report contains the impression description, usually in one sentence. Findings are a description. Tags are a list of keywords. Generating such disparate information from a unified framework is technically demanding.

Secondly, it is still difficult to locate the lesion area in the image and attach the correct description.

Finally, descriptions in imaging reports are often long, containing multiple sentences or even paragraphs. y has S sentences, the i_{th} sentence has N words, and y(i,j) is the j_{th} word in the i_{th} sentence. The loss ℓ(x, y) in long sentences produced by producing distribution values on each word of each sentence consists of two weighted and intersecting terms and a sentence loss ℓ shifts the distribution values when stopped, and the word loss ℓ on the word distribution p(i,j).

However, it is indispensable to generate long texts, and this traditional method cannot meet the needs of long texts.

2.2. Main technologies

Both CNN and RNN are extensions of traditional neural networks, which can generate results by forward calculation, and update the model by reverse calculation. Each layer of neural network can have multiple neurons horizontally, and there can be multiple layers of neural network connections vertically. The significance of the combination is that the combination can process a large amount of information and has the characteristics of time and space, such as video, image and text combination. There are also real scene dialogues and dialogues with images to make text expressions more specific, and videos are more complete than pictures description.

Feature extraction mainly adopts convolution kernel, whose width and height are greater than 1, and which only performs cross-correlation operation with each position of the same size in the image. Therefore, the output size is equal to the input size $n_h \times n_w$ minus the convolution kernel size $k_h \times k_w$, which is:

$$\begin{aligned} \ell(x,y) \quad &= \lambda_{\text{sent}} \sum_{i=1}^{S} \ell_{\text{sent}}\left(p_i, \boldsymbol{I}[i=S]\right) \\ &+\lambda_{\text{word}} \sum_{i=1}^{S} \sum_{j=1}^{N_i} \ell_{\text{word}}\left(p_{ij}, y_{ij}\right) \end{aligned} \tag{2}$$

Image description technology can automatically generate text descriptions for a given image. Most of the image text models studied recently are based on CNN-RNN framework. Vinyals et al. [6] provided image features extracted from the last hidden layer of CNN to LSTM network to generate text. Fang et al. [7] first used CNN to detect anomalies in the image which were used to generate a

complete sentence through the language model. Karpathy et al. [8] put forward the use of multimodal recursive neural network to fuse visual and semantic features and then generate image description.

Scientists have been devoted to studying the attention in the field of cognitive neuroscience since the 19th century. Kernel regression [9] in 1964 was a simple demonstration of machine learning with attention mechanism. Described in mathematical language, suppose there is a query q ∈ Rq and m key-value pairs (k_1, v_1)...,(k_m,v_m), where ki ∈ R_k, v_i ∈ R (v). The attention convergence function F is expressed as a weighted sum of values:

$$\boldsymbol{h}_i = f\left(\boldsymbol{W}_i^{(q)}\boldsymbol{q}, \boldsymbol{W}_i^{(k)}\boldsymbol{k}, \boldsymbol{W}_i^{(v)}\boldsymbol{v}\right) \in \mathbb{R}^{p_v} \tag{3}$$

The attention weight (scalar) of the query q and the key ki is obtained by mapping the two vectors into scalars through the attention scoring function a, and then through the softmax operation:

$$\boldsymbol{W}_o \begin{bmatrix} \boldsymbol{h}_1 \\ \vdots \\ \boldsymbol{h}_h \end{bmatrix} \in \mathbb{R}^{p_o} \tag{4}$$

Attention mechanisms have proven useful for adding image text. Xu et al. introduced spatial visual attention mechanism into image features extracted from CNN middle layer [10]. Wang et al. [11] proposed a semantic attention mechanism for given image tags. In order to make better use of visual features and generate semantic labels.

The design of LSTM network was inspired by the logic gates of computers. LSTM introduces memory cells, or cells for short, whose hidden layer outputs include hidden states and memory elements. Only the hidden state is passed to the output layer, while the memory element is entirely internal information. Suppose there are h hidden units, the batch size is n, and the input number is d. Therefore, the input is X ∈ R (n × d), and the hidden state of the previous time step is H (t − 1) ∈ R (n × h). Accordingly, the gate of time step t is defined as follows: the input gate is I_t ∈ Rn × h, the forgetting gate is F_t ∈ Rn × h, and the output gate is O_t ∈ Rn × h. They are calculated as follows:

$$\begin{aligned} \boldsymbol{I}_t &= \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xi} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hi} + \boldsymbol{b}_i) \\ \boldsymbol{F}_t &= \sigma\left(\boldsymbol{X}_t\boldsymbol{W}_{xf} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hf} + \boldsymbol{b}_f\right) \\ \boldsymbol{O}_t &= \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xo} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{ho} + \boldsymbol{b}_o) \end{aligned} \tag{5}$$

where W_{xi}, W_{xf}, W_{xo} ∈ R (d * h) W_{hi}, W_{hi}, W_{ho} ∈ Rh is the weight parameter, b_i, b_f, b_o ∈ R (l * h) is offset parameters.

As an improved recurrent neural network, LSTM can solve the problem of the long-distance dependence in the process of medical report generation which RNN cannot deal with [12]. Tong et al. [13] are studying intensive text, requiring the model to generate a text description for each detected image region. Lei et al. [14] generated paragraph descriptions for images through layered LSTM.

3. RCLN model

3.1. Model definition

The visual features of the image and the semantic features of the previous sentence are combined

into a multimodal cyclic generation network model (MRNA) that generates the next sentence. The RCLN model proposed in this paper proposes a new cyclic generation model to generate results sentence by sentence, in which subsequent sentences are conditional on multi-modal input, including the preceding sentence and the original sentence image [15]. The multimodal model proposed in this paper adopts attention mechanism to improve performance. The overall architecture presented in this paper takes medical images as input from multiple views and generates a framework for radiology reports with impressions and findings. To generate the survey result paragraphs, this paper first uses an encoder-decoder model, which takes image pairs as inputs and generates the first sentence. The first sentence is then input into the sentence coding network to output the semantic representation of the sentence [16]. Suppose a result paragraph containing L sentences is being generated. The probability of generating the i_{th} sentence of length T satisfies:

$$\mathbb{P}(S_i = w_1, w_2, \ldots, w_T \mid V; \theta)$$
$$= \mathbb{P}(S_1 \mid V)\prod_{j=2}^{i-1}\mathbb{P}(S_j \mid V, S_1, \ldots S_{j-1})\mathbb{P}(w_1 \mid V, S_{i-1})\prod_{t=2}^{T}\mathbb{P}(w_t \mid V, S_{i-1}, w_1, \ldots w_{t-1}) \quad (6)$$

where V is the given medical image, θ is the model parameter (θ on the right is omitted in this paper), Si represents the i_{th} sentence, w_t is the t_{th} mark in the i_{th} sentence. Similar to the n-gram hypothesis in the language model, this paper adopts Markov hypothesis to generate the 2-gram model at sentence level, which means the current sentence being generated depends only on its previous sentence and image. This simplifies the steps to estimate the probability:

$$\hat{\mathbb{P}}(S_i = w_1, w_2, \ldots w_T \mid V; \theta) = \underbrace{\mathbb{P}(S_1 \mid V)}_{1}\underbrace{\prod_{j=2}^{i-1}\mathbb{P}(S_j \mid V, S_{j-1})}_{2}\underbrace{\mathbb{P}(w_1 \mid V, S_{i-1})\prod_{t=2}^{T}\mathbb{P}(w_t \mid V, S_{i-1}, w_1, \ldots w_{t-1})}_{3} \quad (7)$$

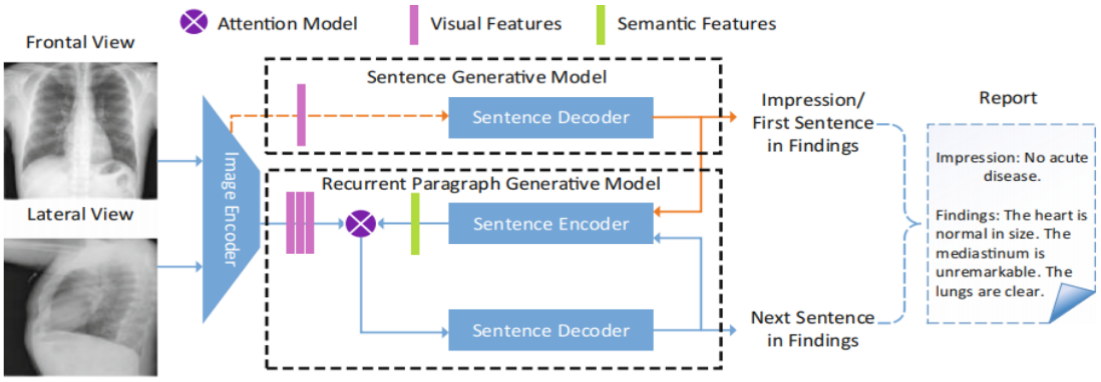

Figure 1. RCLN model flowchart.

It can be noted that for small-scale data sets, the verbal information of medical reports is more important than the image information, and the final results tend to care more about the degree of imitation of doctors' tone.

3.2. Image encoder

The medical reporting task is easily related to the Image2Text task, so this paper utilizes the Image Captions method to solve the problem of this task. In this model, an image encoder is applied to extract global and regional visual features from the input image. The background variable C output

by the image encoder encodes the information of the entire image input sequence $x_1,\ldots,x_T$. Given the output sequence $y_1, y_2,\ldots,y_{T'}$ in the training samples, for each time step t′, the conditional probability of output yt of the image decoder will be based on the previous output sequence y1,…,yt′−1 and the background variable c, which is P(yt′ ∣ y1,…,yt′−1,c).At this time, another cyclic neural network can be used as the decoder to output the time step t 'of the sequence. The decoder takes the output $y_{t'-1}$ of the previous time step and background variable c as the input, and transforms them with the hidden state $s_{t'-1}$ of the previous time step into the hidden state $s_{t'}$ of the current time step. Therefore, function g (cyclic neural network unit) can be used to express the transformation of the hidden layer of the image decoder:

$$st' = g(yt' - 1, c, st' - 1) \tag{8}$$

Image encoders automatically extract visual features of hierarchical CNN images. The image encoder of this model uses pre-trained Resnet-152 [10]. In this paper, the size of the input image is adjusted to 224 × 224 to keep consistent with the image of pre-trained Resnet encoder. Then, the local eigenmatrix f ∈ R1024 × 19 (reconstructed from 1024 × 14 × 14) res layer of Resnet [17]. Each column of f is a regional eigenvector. So, each image has 196 subregions. At the same time, this paper extracts the global feature vector f ∈ R2048 from the last mean pooling layer of Resnet. For multiple input images from multiple views (for example, the front and side views shown in the body text), their regional and global features are connected accordingly before feeding into the following layers [18]. For efficiency, all parameters in the layer built from Resnet-152 are fixed during training. Then, the maximum pooling operation is applied to the feature maps extracted from each convolution layer to generate 1024-dimension feature vectors. The final sentence feature is a concatenation of feature vectors from different layers. To generate a long paragraph description, a hierarchical cycle network was chosen in this paper. A two-level RNN is generally used for paragraph generation: first, some topics are generated by paragraph-level RNN which are then taken as input by a sentence-level RNN to generate sentences. The pre-trained dense subtitle model can be used to detect the semantic regions of images.

3.3. Sentence generation model

Natural language is a complex system used to express the human mind. In this system, words are the basic units of meaning. As the name suggests, a word vector is a vector used to represent the meaning of a word, and can also be considered a feature vector or representation of a word. The technique of mapping words to real vectors is called word embedding. In recent years, word embedding has gradually become the basic knowledge of natural language processing. Word vector is used to represent the word meaning which can also be regarded as the word feature vector. Each word is mapped to a fixed-length vector that better expresses similarities and analogies between different words. Word embedding consists of two models, namely skip-gram and continuous bag of words. For semantically meaningful representations, their training relies on conditional probability, which can be seen as the use of some words in a corpus to predict other words [19]. Word embedding models are self-supervised models since it is unlabeled data.

3.4. Cyclic paragraph generation model

For the Impression and Findings description of medical reports, QA + Hierarchical RNN method was used in this paper to solve this problem [20]. By introducing hidden state variables to store past

information and current input, current output can be determined. Hidden state is a kind of modeling of the way data is generated. It considers that data generation is divided into two steps: first, select a hidden state and then generate observation results from the hidden state [21]. Hiding means you can only see the observation sequence and not the hidden state sequence when the data generation is on the run, but it doesn't affect the hidden state being exposed to you during training [22]. All of this is done in the basic unit of time step, and the time step is the time interval of the load sub-step in the load step [23]. In rate-independent analysis such as static analysis and (static) nonlinear analysis, in a load step, the time step does not reflect the real time, it is accumulated to reflect the sequence of load sub-steps [24]. However, in rate-dependent analysis such as transient analysis, the size of time step reflects actual length of time.

4. Experiment

4.1. Dataset

The original dataset was collected from Openi's chest radiography open data, which contained 3955 radiology reports from two large hospital systems in the Indiana Patient Care Network database and 7470 related chest X-rays from the Hospital Image Archiving System.

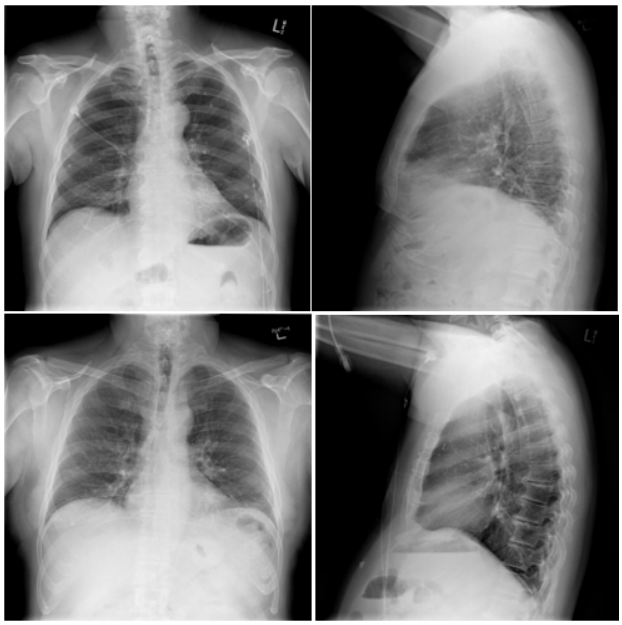

Figure 2. Sample dataset picture.

First, the original data set contained 7470 images, 3391 pairs of positive side chest radiographs and 3631 pairs of sentences of which the number of sentences is greater than 4. In order to ensure that the largest subset of data information can be obtained, the maximum number of sentences was set to 8 since more than 90% of report statements are between 4 and 8 sentences. There were 3111 applications that met both conditions. Secondly, the training and validation dataset are spilt into 2811/300 with a ratio of about 1/10, using Adam optimization function based on stochastic gradient descent. The unused part of the dataset is then used as the test set. In this paper, 300 reports were randomly selected to form a test set on which all evaluations were performed.

4.2. Evaluation indicator

Some common image caption evaluation metrics, including bilingual evaluation understudy (BLEU), metric for evaluation of translation with explicit ordering (METEOR), and recall-oriented understudy for gisting evaluation (ROUGE), are used to provide quantitative comparisons in this paper. BlEU-1 measures the accuracy of words in medical reports, and higher-order BlEU can measure the fluency of sentences. For a sentence to be translated, candidate translations can be expressed as, and the corresponding group of reference translations can represent the phrase set of n words and the possible grams of the kth group.

The purpose of METEOR is to prevent mistranslations of the reported results due to synonyms [25]. The measurement of METEOR is based on weighted harmonic mean value of single precision and single word recall rate. To calculate METEOR, a set of alignments needs to be given in advance, which is based on the thesaurus of WordNet. The alignments are calculated as harmonic average of accuracy and recall rate between the corresponding best candidate translation and reference translation the METEOR is calculated as the harmonic mean of precision and recall rate between corresponding best candidate translation and reference translations by minimizing successive ordered chunks in the corresponding statement:

$$\mathrm{Pen} = \gamma\left(\frac{ch}{m}\right)^{\theta} \tag{9}$$

$$F_{\mathrm{mean}} = \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m} \tag{10}$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \tag{11}$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \tag{12}$$

$$METEOR = (1 - Pen) F_{\mathrm{mean}} \tag{13}$$

where α, γ and θ are the default parameters for evaluation. Therefore, the final evaluation of METEOR is based on a harmonic average of decomposition matching and characterization decomposition matching quality of chunk, and contains a penalty coefficient Pen which is different from BLEU. Accuracy and recall rate based on the whole corpus are taken into account to obtain the final measure.

ROUGE evaluates abstracts based on the co-occurrence information of n-grams in abstracts, and it is a method for evaluating the recall rate of n-gram words based on the co-occurrence information of n-gram words [26]. The basic idea is that several experts generate artificial abstracts respectively to form a standard abstract set. The quality of the abstract is evaluated by counting the number of overlapping basic units (n-element grammar, word sequences and word pairs) through comparing the automatic abstracts generated by the system with the standard abstracts generated by the manual.

$$\text{ROUGE-N}= \frac{\sum_{s\in\{\text{ReferemceSumaaries}\}} \sum_{\text{gramn}} \text{Count}_{\text{match}}(\text{gramn})}{\sum_{s\in\{\text{ReferenceSummaries}\}} \sum_{\text{gramn}\in s} \text{Count}(\text{gramn})} \tag{14}$$

The stability and robustness of the evaluation system can be improved by comparing with the expert manual abstract. Neural machine translation (NMT) used in this paper is more powerful than its predecessor Statistical machine translation (SMT). The word order of medical reports is often correct but the error frequency increases. Therefore, a recall rate indicator like ROUGE is needed to evaluate the error frequency.

4.3. Contrast test

Firstly, an image encoder is used to extract global and regional visual features from the input image. Image encoder is a CNN, which automatically extracts hierarchical visual features from images. More specifically, we adjust the size of the input image to 224 × 224. (Corresponding to the image size parameter).

As shown in Figures 3–5, a dropout layer (corresponding to the dropout rate parameter) with a value of [0.3, 0.5 or 0.7] has been added to the network to reduce overfitting and this dropout layer represents the probability that the layer's output is discarded.

Figure 3. dropout = 0.3.

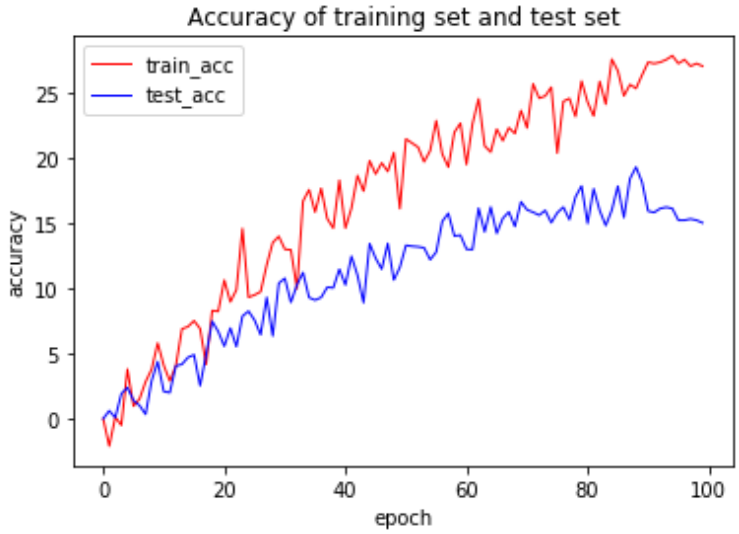

Figure 4. dropout = 0.5.

Word embedding is mainly responsible for processing the title of each image given as input during training. The output of the word embedding is also a vector of size 1 × 256 (corresponding to the argument word_embedding_size parameter), which is another input to the decoder sequence.

Start the training, set the batch size to 32 (corresponding to the parameter batch size), Adam optimizer makes the learning rate from 1E-2 to 1E-4 (corresponding to the parameter learning rate), a total of 50 iterations (parameter epoch num).

The probability and accuracy influence of network layer output being discarded are discussed.

Figure 5. dropout = 0.7.

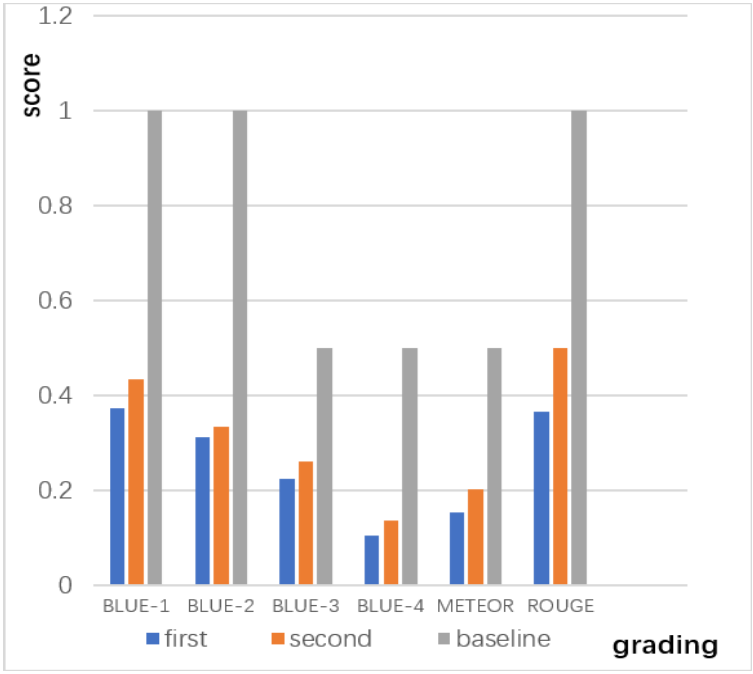

Figure 6. Comparison between RCLN model data and reference data in two experiments; The horizontal axis represents different score names, and the vertical axis represents the score value.

In the following two model tests, the data output of the first and second tests both met the evaluation benchmark range. The label position of the model was adjusted before the second test. The

performance of various indicators was improved when the label at the end of the whole sentence was changed to the half of the sentence and the training time was increased. The time complexity of this model is O(n^2). As shown in Figures 6 and 7, the minimum values of the baseline range are all 0. It can be seen that each score index is lower than the maximum value, proving that this model can generate relatively standard medical reports.

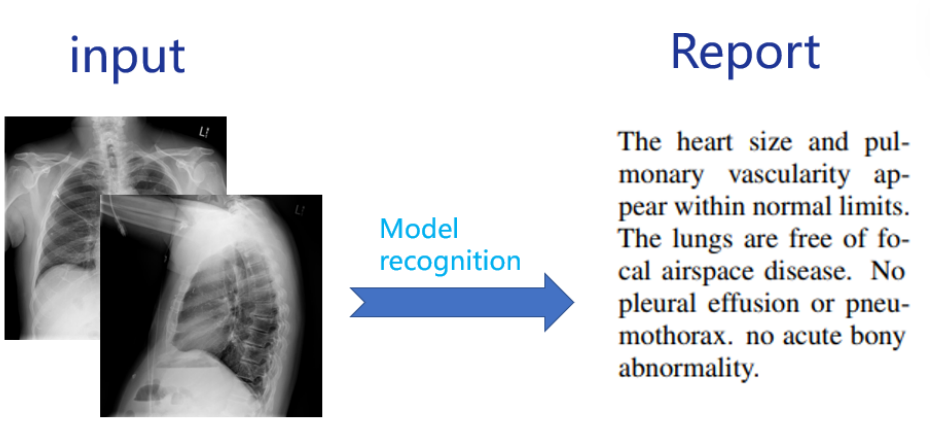

Figure 7. Model input and output test examples.

In this paper, two comparative models for medical report generation are also implemented. The same Resnet pre-training model was used for pre-training. The data results are shown in Table 1.

CNN-NN, the prototype CNN, was published by Lecun in 1998 [27]. He formally proposed that he applied the back propagation to neural networks and proposed a new neural network convolution NN. Ronald Williams and David Zipser put forward real-time circular learning of RNN as the basis in 1989 [28].

CNN-RNN-Att: The Attention mechanism was added on the basis of the previous one. The Attention mechanism was published by google mind team in 2014 [29]. In 2017, the article "Attention is All You Need" was published by Google Machine Translation team in which self-attention mechanism was extensively used to learn text representation.

Table 1. Comparison of the model.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE
CNN-RNN	0.3063	0.2026	0.148	0.0994	0.1525	0.3273
CNN-RNN-Att	0.3235	0.2374	0.1197	0.1084	0.1484	0.3256
MRNA	0.3773	0.2436	0.1726	0.1284	0.1635	0.3263
RCLN	0.4341	0.3336	0.2623	0.1373	0.2034	0.3663

By comparing the results of other models and RCLN models, it can be seen that the model based on the multi-attention mechanism is superior to similar models in terms of BLUEs, METEOR and ROUGE, indicating the effectiveness of multi-attention mechanism on medical report generation [30]. The scores of RCLN model were much higher than CNN-RNN series model and higher than MRNA model, proving its effectiveness. Some statements in reports generated by other models are

continuous but not coherent. In contrast, the model proposed in this paper is more coherent in context and more colloquial.

5. Conclusions

This paper mainly focusses on generating detailed findings for chest radiographs medical reports. For impression generation, classification-based methods may be better at distinguishing anomalies and then drawing final conclusions. But from the results, we can see that in the first line the results and impressions are consistent with the actual situation. However, the results and impressions generated in the second line leave out some exception descriptions. The main reason may be that I was training on a small training set, with fewer training samples for anomalies, and some inconsistencies caused by real noise from the original report. Furthermore, the current model does not create high-quality new sentences that never appear in the training set. The reason may be that it is difficult to learn correct grammar from a small corpus because syntactic correctness is not considered in the training objective function.

In conclusion, it is believed that with more control data sets and better noise reduction processing of data set preprocessing, better results will appear [31]. At the same time, multiple loop processing statements can also increase the depth, making the result more accurate. In the data labeling process, the addition of more high-quality sentences is expected to effectively ensure the enhancement of the quality of the results.

Acknowledgments

The research is supported by the National Natural Science Foundation of China (No.12105120, No.72174079, No.72101045), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No.19KJB520004，No.21KJB520033), Jiangsu Province “333” project (BRA2020261), Jiangsu Qinglan Project, Lianyungang “521 project”, Science and Technology project of Lianyungang High-tech Zone (No.ZD201912).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Q. Z. You, H. L. Jin, Z. W. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 4651–4659. http://doi.org/10.1109/CVPR.2016.336
2. J. Krause, J. Johnson, R. Krishna, F. F. Li, A hierarchical approach for generating descriptive image paragraphs, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3337–3346. http://doi.org/10.1109/CVPR.2017.356
3. R. K. Meleppat, L. K. Seah, M. V. Matham, Spectral phase-based automatic calibration scheme for swept source-based optical coherence tomography systems, *Phys. Med. Biol.*, **61** (2016), 7652–7663. https://doi.org/10.1117/12.2190530

4. R. K. Meleppat, *In vivo* multimodal retinal imaging of disease-related pigmentary changes in retinal pigment epithelium, *Sci. Rep.*, **11** (2021), 1–14. https://doi.org/10.1088/0031-9155/61/21/7652
5. R. K. Meleppat, Plasmon resonant silica-coated silver nanoplates as contrast agents for optical coherence tomography, *J. Biomed. Nanotechnol.*, **12** (2016), 1929–1937. https://doi.org/10.1166/jbn.2016.2297
6. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput*., **9** (1997), 1735–1780. http://doi:10.1162/neco.1997.9.8.1735
7. H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, et al., From captions to visual concepts and back, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2015), 1473–1482. http://doi.org/10.1109/CVPR.2015.7298754
8. K. Andrej, F. F. Li, Deep visual semantic alignments for generating image descriptions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3128–3137. http://doi.org/10.1109/TPAMI.2016.2598339
9. H. Bierens, The Nadaraya-Watson kernel regression function estimator, in *Topics in Advanced Econometrics*: *Estimation*, *Testing*, *and Specification of Cross-Section and Time Series Models*, (1994), 212–247. http://doi.org/10.1017/CBO9780511599279.011
10. X. Wang, Z. Duan, L. Liu, M. Li, Y. An, Y. Zhou, Multi-Timescale load forecast of large power customers based on online data recovery and time series neural networks, *J. Circuits Syst. Comput.*, **31** (2022), 2250088. http://doi.org/10.1142/S0218126622500888
11. S. Wang, X. Ye, Y. Gu, J. Wang, Y. Meng, J. Tian, et al., Multi-label semantic feature fusion for remote sensing image captioning, *ISPRS J. Photogramm. Remote Sens.*, **2022** (2022), 1–18. http://doi.org/10.1016/j.isprsjprs.2021.11.020
12. F. Christophe, Learning algorithm recommendation framework for IS and CPS security: Analysis of the RNN, LSTM, and GRU contributions, *Int. J. Syst. Software Secur. Prot.*, **13** (2022), 1–8. http://doi.org/10.4018/IJSSSP.293236
13. G. Tong, Y. Li, D. Chen, Q. Sun, W. Cao, G. Xiang, CSPC-Dataset: New LiDAR point cloud dataset and benchmark for large-scale semantic segmentation, *IEEE Access*, **8** (2020), 87695–87718. http://doi.org/10.1109/ACCESS.2020.2992612
14. J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, M. Bansal, MART: Memory-augmented recurrent transformer for coherent video paragraph captioning, preprint, arXiv:2005.05402.
15. Z. F. Li, Y. Q. Yang, L. P. Wu, Study of text sentiment analysis method based on GA-CNN-LSTM model, *J. Jiangsu Ocean Univ. (Nat. Sci. Ed.)*, **30** (2021), 79–86.
16. H. Li, X. P. Ma, J. Shi, C. Li, Z. Zhong, H. Cai, A recommendation model by means of trust transition in complex network environment, *Acta Autom. Sin.*, **44** (2018), 363–376. http://doi.org/10.16383/j.aas.2018.c160395
17. Y. Ma, P. Feng, P. He, Y. Ren, X. Guo, X. Yu, et al., Segmenting lung lesions of COVID-19 from CT images via pyramid pooling improved Unet, *Biomed. Phys. Eng. Express*, **7** (2021), 45008. http://doi.org/10.1088/2057-1976/ac008a
18. H. Y. Chung, Automatische evaluation der Humanübersetzung: BLEU vs. METEOR, *Lebende Sprachen*, **65** (2020), 25–36. http://doi.org/10.1515/les-2020-0009
19. C. Zhao, Y. Xu, Z. He, J. Tang, Y. Zhang, J. Han, et al., A new approach for lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images, *Pattern Recognit.*, **119** (2021), 108071–108079. https://doi.org/10.1016/j.patcog.2021.108071
20. S. A. Thorat, K. P. Jadhav, Improving conversation modelling using attention based variational hierarchical RNN, *Int. J. Comput.*, **20** (2021), 39–45. http://doi.org/10.47839/ijc.20.1.2090

21. H. M. Sabbir, Att-BiL-SL: Attention-based Bi-LSTM and sequential LSTM for describing video in the textual formation, *Appl. Sci.*, **12** (2021), 1–8. http://doi.org/10.3390/app12010317
22. N. Mu, H. Y. Wang, Y. Zhang, J. Jiang, J. Tang, Progressive global perception and local polishing network for lung infection segmentation of COVID-19 CT images, *Pattern Recognit.*, **120** (2021), 108168. https://doi.org/10.1016/j.patcog.2021.108168
23. X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, et al., Weakly supervised segmentation of COVID-19 Infection with scribble annotation on CT images, *Pattern Recognit.*, **122** (2022), 108341–108349. https://doi.org/10.1016/j.patcog.2021.108341
24. J. He, Q. Zhu, K. Zhang, P. Yu, J. Tang, An evolvable adversarial network with gradient penalty for COVID-19 infection segmentation, *Appl. Soft Comput.*, **113** (2021), 107947–107956. https://doi.org/10.1016/j.asoc.2021.107947
25. D. Deutsch, T. B Weiss, D. Roth, Towards question-answering as an automatic metric for evaluating the content quality of a summary, *Trans. Assoc. Comput. Linguist.*, **9** (2021), 774–789. http://doi.org/10.1162/TACL_A_00397
26. F. P Martin, H. Weishaar, F. Cristea, J. Hanefeld, L. Schaade, C. E. Bcheraoui, Impact of type and timeliness of public health policies on COVID-19 epidemic growth: Organization for economic co-operation and development (OECD) member states, January–July 2020, *SSRN Electron. J.*, **2020** (2020), 1–8. http://doi.org/10.2139/ssrn.3698853
27. Y. Lecun, L. Bottou, Gradient-based learning applied to document recognition, *Proc. IEEE*, **86** (1998), 2278–2324. http://doi.org/10.1109/5.726791
28. R. J. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, *Neural Comput.*, **1** (1989), 270–280. https://doi.org/10.1162/neco.1989.1.2.270
29. D. Buchan, D. T. Jones, Learning a functional grammar of protein domains using natural language word embedding techniques, *Proteins Struct. Funct. Bioinf.*, **88** (2020), 2555. http://doi.org/10.1002/prot.25842
30. D. Jia, Y. Fujishita, C. Li, Y. Todo, H. Dai, Validation of large-scale classification problem in dendritic neuron model using particle antagonism mechanism, *Electronics*, **9** (2020), 792. http://doi.org/10.3390/electronics9050792
31. X. G. Lv, X. M. Sun, G. L. Zhu, L. Jiang, S. T. Lu, Research on image smoothing and texture extraction based on variational method, *J. Jiangsu Ocean Univ. (Nat. Sci. Ed.)*, **30** (2021), 77–84.

AIMS Press