Retinal OCT image report generation based on visual and semantic topic attention model

Guo, Chao, Zhu, Weifang, Wang, Ting, Lin, Tian, Chen, Haoyu, et al.

Chao Guo, Weifang Zhu, Ting Wang, Tian Lin, Haoyu Chen, Xinjian Chen, "Retinal OCT image report generation based on visual and semantic topic attention model," Proc. SPIE 12032, Medical Imaging 2022: Image Processing, 120322C (4 April 2022); doi: 10.1117/12.2611469

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Retinal OCT Image Report Generation Based on Visual and Semantic Topic Attention Model

Chao Guo[1], Weifang Zhu[1], Ting Wang[1], Tian Lin[3], Haoyu Chen[3], Xinjian Chen[1,2,*]
[1]School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu Province, 215006, China
[2]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, Jiangsu Province, 215123, China
[3]Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong, Shantou, Guangdong Province, 515000, China

ABSTRACT

Optical coherence tomography (OCT) is widely used in the diagnosis of retinal diseases. Reading OCT images and summarizing its insights is a routine, yet nonetheless time-consuming task. Automatic report generation can alleviate this issue. There are two major challenges in this task: (1) An OCT image may contain several fundus abnormalities and it is difficult to detect them all simultaneously. (2) The diagnostic reports are complex, which need to describe multiple lesions. In this paper, we propose a deep learning-based model, named as VSTA model (Visual and Semantic Topic Attention model), which is able to generate report from the input OCT image. Our major contributions include: (1) Semantic attention and visual attention are jointly embedded to the model to generate diagnosis report with complex content. (2) Semantic tags based on image similarity is employed to initialize the semantic attention weights, which increases the prediction accuracy of the model. With the proposed VSTA model, the metric of BLEU-4, CIDEr and ROUGE-L reach 31.16, 264.22 and 52.58, which are better than some existing advanced methods.

Keywords: Optical coherence tomography, automatic report generation, visual and semantic attention mechanism

1. INTRODUCTION

Optical coherence tomography (OCT) is widely used in the diagnosis and treatment of ophthalmic diseases because of its non-invasive and high resolution [1]. The ophthalmic diagnostic report, shown in Figure 1, serves as an interpretation of the medical image and describes the findings of each retinal region, whether it is abnormal or potentially abnormal. The reading and interpretation of medical image are usually conducted by specialized medical professionals. Due to the large number of patients, it costs a lot of time of the experts to write reports. In regions with backward medical treatment, there is a lack of professional physicians to write diagnostic reports. The computer-assisted OCT medical report generation can greatly reduce the workload of ophthalmologists and help them make decisions.

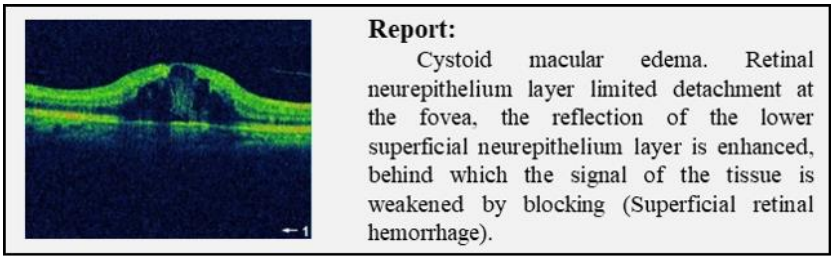
Automatic OCT image report generation can be viewed as a task of image captioning. This task contains two major challenges. First, an OCT image may contain several retinal abnormalities and detecting all the abnormal regions simultaneously is challenging. Second, the content of the diagnostic report is very complex. It contains the description of all abnormalities on the patient's retina, which increases the difficulty of report generation. Figure 1 illustrates a diagnostic report, which describes different kinds of diseases such as edema, abnormality of neurepithelium layer structure, abnormality of neurepithelium layer reflection, etc. Besides, each type of diseases has its own terminology and description, for example, the description of edema includes: diffuse edema, cystoid edema, macular edema and so on.

In recent years, there have been several related studies on medical image captioning [2,3,4,5]. Jing et al. used a hierarchical LSTM model with co-attention mechanism to generate diagnostic reports of chest x-ray images [2]. Liu et.al. proposed a Posterior-and-Prior Knowledge Exploring-and-Distilling approach to imitate the working patterns of radiologists for report generation [3]. These studies are for x-ray images or ultrasound images, but there are few studies on ophthalmic image report generation. In this paper, a visual and semantic topic attention combined model, named as VSTA model, is proposed to generate accurate and detailed retinal OCT image diagnostic reports.

*Corresponding author: E-mail: xjchen@suda.edu.cn.

Medical Imaging 2022: Image Processing, edited by Olivier Colliot,
Ivana Išgum, Proc. of SPIE Vol. 12032, 120322C · © 2022 SPIE
1605-7422 · doi: 10.1117/12.2611469

In our work, each disease is referred to as a semantic topic and diagnostic reports consists of sentences with different semantic topics. For generating such reports, a semantic topic embedding matrix is added in the model. At each time step, semantic attention will guide the model to generate sentences with different semantic topics. At the same time, the visual attention will guide the model to focus on different regions of image, to generate the specific content of the sentence. In addition, semantic tags, based on image similarity, are employed to initialize the semantic attention weights, increasing the prediction accuracy of the model.


Report:

Cystoid macular edema. Retinal neurepithelium layer limited detachment at the fovea, the reflection of the lower superficial neurepithelium layer is enhanced, behind which the signal of the tissue is weakened by blocking (Superficial retinal hemorrhage).

Figure 1. Retinal OCT image and report.

2. METHODS

In this section, the proposed method is described as four parts: network structure of Visual and Semantic Topic Attention model, semantic attention and topic embedding, semantic attention initialization and visual attention.

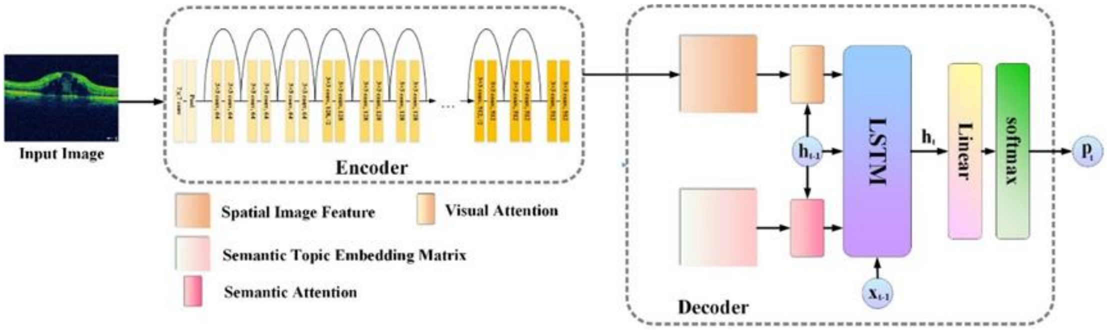


Figure 2. VSTA model structure.

2.1 Network structure of Visual and Semantic Topic Attention model

As shown in Figure 2, VSTA model is an encoder-decoder structure. In the part of image encoder, a Resnet101 [6] without the last fully connected layer is used to extract the features $V \in \mathbb{R}^{m \times d}$ of different regions of the input image I. The decoder consists of Long-Short Term Memory (LSTM) network [7], visual attention φ and semantic attention ϕ. The main working flow of VSTA model is governed by the following equations:

$$V = CNN(I) \tag{1}$$

$$c_t = \varphi(V, h_{t-1}), \quad t > 0 \tag{2}$$

$$\hat{s}_t = \phi(h_{t-1}), \quad t > 0 \tag{3}$$

$$h_t = LSTM(h_{t-1}, c_t, \hat{s}_t, x_{t-1}) \tag{4}$$

$$p_t = soft\max(W_p(h_t)) \tag{5}$$

Where c_t, $\hat{s}_t$ are context vectors generated by visual and semantic attention respectively, x_{t-1} is the word embedding of step *t-1*, W_p is the weight parameters to be learnt. Eq. (2) to (4) are recursively applied to generate hidden state h_t. Finally, the probability p_t over a vocabulary of possible words at time *t* can be calculated by Eq. (5).

2.2 Semantic attention and topic embedding

Inspired by word embedding, the semantic topics contained in the corpus are abstracted as *k* learnable embeddings $s \in \mathbb{R}^{1\times d}$. These *k* embeddings form a topic embedding matrix $S = [s_1, s_2, \cdots, s_k], S \in \mathbb{R}^{k\times d}$. During the training, the model can autonomously learn the embedding representation of semantic topics and the correlation between topics. A semantic attention ϕ is proposed to compute the context topic vector $\hat{s}_t$, which is defined as Eq. (3). The hidden state h_{t-1} at step *t-1* is fed in a signal linear layer followed by a softmax function to generate the attention distribution $\alpha_{s,t}$ over the *k* semantic topic embedding:

$$\alpha_{s,t} = soft\max(W_{s,h}(h_{t-1})) \tag{6}$$

where $W_{s,h}$ is parameters to be leant. Based on the attention distribution $\alpha_{s,t}$, the context topic vector can be obtained by:

$$\hat{s}_t = \sum_{i=1}^{k} \alpha_{s,t,i} s_i \tag{7}$$

where $\alpha_{s,t,i}$ is the *i'th* element in $\alpha_{s,t}$, s_i is the *i'th* topic embedding vector in topic embedding matrix *S*.

2.3 Semantic attention initialization

Since OCT images often contain many diseases, in order to accurately predict the reports, the semantic tags, which is generated by image similarity, is encoded into a one-hot vector as the initial semantic attention weights. The details are shown in Figure 3. Firstly, the image local features are extracted by Scale-Invariant Feature Transform (SIFT) [8]. Secondly, the local features of all the images in the training set were collected, and k-means was used for clustering to obtain k clustering centers, each of which corresponds to a semantic topic. After that, the number of local features, $N = [n_1, n_2, \cdots, n_k]$, contained by each semantic topic in the image can be obtained. Finally, a pre-set threshold *th* is used to generate initial semantic attention distribution $\alpha_{s,0}$ as following:

$$\alpha_{s,0,i} = \begin{cases} 0, & n_i < th \\ 1, & others \end{cases}, \quad i = 0, \ldots, k \tag{8}$$

where $\alpha_{s,0,i}$ is the *i'th* element in $\alpha_{s,0}$, n_i is the *i'th* element in *N*.

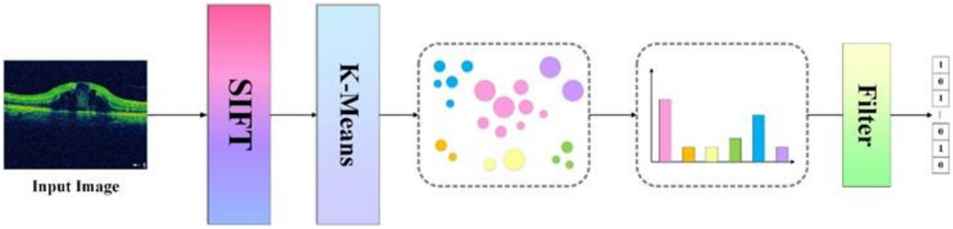


Figure 3. Initialization for semantic attention.

2.4 Visual attention

To generate the specific content of the sentence, the visual attention φ is employed to compute the context vector c_t, making model focus on different regions of image *I*. The context vector is defined as Eq. (9) and (11). Given the spatial image feature $V = [v_1, v_2, \cdots, v_m]$, and hidden state h_{t-1}, the context vector c_t is generated as:

$$z_{v,t} = \mathrm{Re}LU(W_{v,h}(W_v V + W_h h_{t-1})) \quad (9)$$

$$\alpha_{v,t} = soft\max(z_{v,t}) \quad (10)$$

$$c_t = \sum_{i=1}^{m} \alpha_{v,t,i} v_i \quad (11)$$

where $W_{v,h}$, W_v , W_h are parameters to be leant, $\alpha_{v,t}$ is the visual attention distribution at step *t*.

3. RESULTS

3.1 Datasets and experiment settings

The data used in this paper were provided by Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong. The collection and analysis of image data were approved by the Institutional Review Board of Shantou University and adhered to the tenets of the Declaration of Helsinki. The experimental dataset was drawn from 800 patients. The ground truth is written by senior ophthalmologist. Data from 600 patients was used for training and data from 200 patients was used for testing.

The proposed method is implemented on the publicly available Pytorch platform and OpenCV libraries. The loss for segmentation model is minimized by the Adam optimizer with an initial learning rate of 1e-3. The network is trained on Nvidia GTX1080. In the training process, the training images are resized to 256×256 and the batch size is set to 16.

3.2 Evaluation metrics

The Bilingual Evaluation Understudy (BLEU) [9], Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [10] and Consensus-based Image Description Evaluation (CIDEr) [11] are employed to quantitatively analyze the experimental results. In particular, BLEU was originally proposed as a machine translation evaluation, with the basic idea of calculating the accuracy of N-grams in candidate sentences. ROUGE-L is designed to measure the quality of the results in automatic summarization task. The idea of ROUGE-L is using the longest common subsequence and F1-score to measure the similarity between the reference text and the candidate text. CIDEr is designed to evaluate the image captioning system. The cosine angle of the term frequency–inverse document frequency (TF-IDF) vectors of the reference text and the candidate text is used to evaluate the similarity between them in CIDEr.

3.3 Results

To evaluate the performance of our method, we compared VSTA model with some advanced models [12,13,14] and set the visual attention model [12] as the baseline. The results are shown in Table 1. Compared with advanced methods, VSTA model achieves better performance on BLEU, ROUGE-L, and CIDEr.

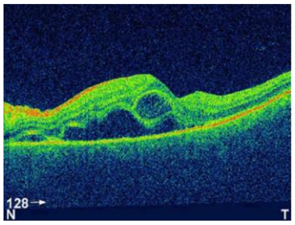
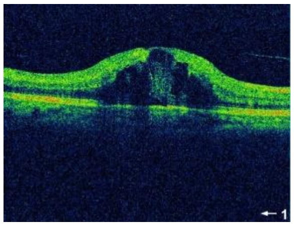
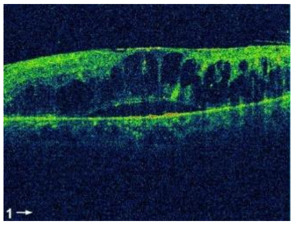
To evaluate the performance of semantic attention initialization mentioned in section 2.3, We randomly initialized semantic attention in the VSTA model and trained it on our data set. The result is shown in Table 1 (Ours-with random initialization). The result shows that the proposed method for semantic attention initialization has a positive effect on the model performance.

In addition, Table 2 shows the reports generated by the baseline and VSTA model. Compared with baseline, VSTA model can identify abnormalities of patients' retina and describe the abnormal areas better. Moreover, the VSTA model achieves better results for diagnostic reports of long length.

Table 1. The values of BLEU, ROUGE-L, CIDEr compared with different methods.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
Visual Attention [12] (baseline)	45.03	37.35	32.46	28.60	233.40	52.53
Visual Sentinel [13]	42.35	34.83	30.88	27.85	236.14	49.03
Bottom-up and top-down [14]	38.65	31.73	28.49	26.60	221.83	45.07
Ours-with random initialization	44.75	37.15	32.99	29.75	252.72	51.77
Ours-with tag initialization	**45.36**	**38.10**	**34.19**	**31.16**	**264.22**	**52.58**

Table 2. Results for some OCT images. (a) Input OCT image. (b) Ground truth. (c) The results of baseline. (d) The results of VSTA model.

	**Report**		
	Retinal edema. Retinal detachment.	Retinal neurepithelium layer edema and limited detachment. The reflection of RPE layer is not uniform.	Retinal edema. Retinal detachment.
	**Report:**		
	Cystoid macular edema. Retinal neurepithelium layer limited detachment at the fovea, the reflection of the lower superficial neurepithelium layer is enhanced, behind which the signal of the tissue is weakened by blocking (Superficial retinal hemorrhage).	Cystoid macular edema. Retinal neurepithelium layer limited detachment at the fovea.	Cystoid macular edema. Retinal neurepithelium layer limited detachment at the fovea, the reflection of the lower superficial neurepithelium layer is enhanced, behind which the signal of the tissue is weakened by blocking (Superficial retinal hemorrhage).
	**Report:**		
	Cystoid macular edema. Local anterior macular membrane. Local retinal neurepithelium layer limited detachment，on which there are scattered high reflex spots, clumps (exudate). The reflection of RPE layer is not uniform.	Cystoid macular edema. Scattered high reflex spots, clumps (exudate) can be seen on the retinal neurepithelium layer. The reflection of RPE layer is not uniform and there are some small, highly reflective projections on the RPE layer.	Cystoid macular edema. Local retinal neurepithelium layer limited detachment，on which there are scattered high reflex spots, clumps (exudate). The reflection of RPE layer is not uniform.
(a)	(b)	(c)	(d)

4. CONCLUSIONS

In this paper, we propose a deep learning based VSTA Model, which can generate diagnostic reports by inputting retinal OCT images. In order to generate complex reports containing multiple topics, we abstract semantic topics included in the corpus into a learnable semantic topic embedding matrix, and adopt visual and semantic attention to generate reports. What's more, we also initialize the semantic attention distribution through the image similarity, which improves the performance of the model. Experiments have proved that the VSTA model achieved good performance on the retinal OCT image report generation.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700, in part by the National Nature Science Foundation of China under Grant U20A20170.

REFERENCES

[1] Abràmoff, et al., "Retinal Imaging and Image Analysis," IEEE Reviews in Biomedical Engineering 3(3), 169–208 (2010).

[2] Jing B, Xie P, Xing E, "On the Automatic Generation of Medical Imaging Reports," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 2577-2586 (2017).

[3] LIU, Fenglin, et al., "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13753-13762 (2021).

[4] Gaurav O. Gajbhiye, Abhijeet V. Nandedkar, et al., "Automatic Report Generation for Chest X-Ray Images: A Multilevel Multi-attention Approach," Computer Vision and Image Processing, 174–182 (2020).

[5] Mohammad Alsharid, Harshita Sharma, et al., "Captioning Ultrasound Images Automatically," Medical Image Computing and Computer Assisted Intervention (MICCAI), 338–346 (2019).

[6] He K, Zhang X, Ren S, et al., "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016).

[7] Hochreiter S, Schmidhuber J, "Long short-term memory," Neural computation 9(8), 1735-1780 (1997).

[8] Lindeberg T, "Scale invariant feature transform," 10491 (2012).

[9] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J, "Bleu: a method for automatic evaluation of machine translation," Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311-318 (2002).

[10] LIN, Chin-Yew, "Rouge: A package for automatic evaluation of summaries," Text summarization branches out, 74-81 (2004).

[11] VEDANTAM, Ramakrishna; LAWRENCE ZITNICK, et al., "Cider: Consensus-based image description evaluation," Proceedings of the IEEE conference on computer vision and pattern recognition 4566-4575 (2015).

[12] Xu K, Ba J, Kiros R, et al., "Show, attend and tell: Neural image caption generation with visual attention," International conference on machine learning, 2048-2057 (2015).

[13] Lu J, Xiong C, Parikh D, et al., "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," Proceedings of the IEEE conference on computer vision and pattern recognition, 375-383 (2017).

[14] Anderson P, He X, Buehler C, et al., "Bottom-up and top-down attention for image captioning and visual question answering," Proceedings of the IEEE conference on computer vision and pattern recognition, 6077-6086 (2018).