Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation

Xiancheng Xie[1], Yun Xiong[1,2(✉)], Philip S. Yu[1,2,3], Kangan Li[4], Suhua Zhang[5], and Yangyong Zhu[1,2]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
{17212010043,yunx}@fudan.edu.cn
[2] Shanghai Institute for Advanced Communication and Data Science, Fudan University, Shanghai, China
[3] Computer Science Department, University of Illinois at Chicago, Chicago, IL 60607, USA
[4] Shanghai General Hospital, Shanghai, China
[5] Department of Nephrology, Suzhou Kowloon Hospital, Shanghai Jiaotong University, School of Medicine, Suzhou 2015028, Jiangsu Province, China

Abstract. Radiology report writing is error-prone, time-consuming and tedious for radiologists. Medical reports are usually dominated by a large number of normal findings, and the abnormal findings are few but more important. Current report generation methods often fail to depict these prominent abnormal findings. In this paper, we propose a model named Attention-based Abnormal-Aware Fusion Network (A3FN). We break down sentence generation into abnormal and normal sentence generation through a high level gate module. We also adopt a topic guide attention mechanism for better capturing visual details and develop a context-aware topic vector for model cross-sentence topic coherence. Experiments on real radiology image datasets demonstrate the effectiveness of our proposed method.

Keywords: Radiology report generation · Attention · Abnormal-aware

1 Introduction

In recent years, radiology images are playing a vital role in the auxiliary diagnosis. Most of the existing captioning methods [3,4,6] perform poorly since report usually consists of multiple long, structural and informative sentences. Besides generating short captions, pioneering radiology report generation method [2] adopts a hierarchical LSTM framework combined with a co-attention mechanism to generate paragraph on IU X-Ray dataset [1]. However, their generated reports tend to describe normal findings with some repetitions and are incapable of capturing rare but prominent abnormality. Xue et al. [5] adopt a feedback

G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 448–452, 2019.
https://doi.org/10.1007/978-3-030-18590-9_64

mechanism to learn long sequence dependency. However, their method is prone to generate fluent but general report without prominent abnormal narratives.

In this paper, we propose an Attention-based Abnormal-Aware Fusion Network (A3FN). Our method can effectively capture these rare but prominent abnormal observations. For more accurate abnormal descriptions, we adopt a topic guide attention mechanism to support abnormal findings with its detailed visual context (e.g., location, size, and severity etc.). We also develop a context-aware topic vector to control topic coherence and descriptive completeness. Our proposed A3FN method achieves the highest detection accuracy of positive abnormality terminologies.

2 The Proposed A3FN Method

This section describes our proposed A3FN in detail. Figure 1 shows the architecture of A3FN. We will go through vital parts of the model in following section.

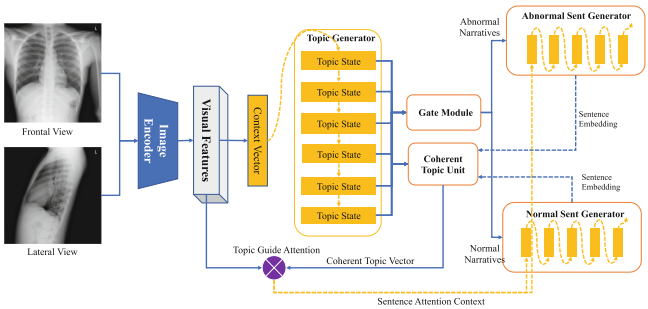

Fig. 1. The architecture of the proposed A3FN method.

Topic Generator and Coherent Topic Unit. Topic generator is a single-layer LSTM which generates a sequence of high-level topic vectors $\mathbf{q}$, one for each sentence. Generally, topic generator can be written as:

$$\boldsymbol{c}_i^t = F_{attn}^t(\boldsymbol{h}^c, \boldsymbol{h}_{i-1}^t), \boldsymbol{h}_i^t = F_{RNN}^t(\boldsymbol{c}_i^t, \boldsymbol{h}_{i-1}^t),$$

$$\boldsymbol{q}_i = tanh(\boldsymbol{W}_t \boldsymbol{h}_i^t + \boldsymbol{b}_t), \boldsymbol{y}_i = softmax(\boldsymbol{W}_{stop} \boldsymbol{h}_i^t + \boldsymbol{b}_{stop}).$$

where F_{attn}^t is the attention mechanism same as [4], and F_{RNN}^t is a single-layer LSTM. The probability $\boldsymbol{y}_i$ controls whether to stop generating next topic vector. After generate all topic states, we enhance current topic state $\boldsymbol{q}_i$ with global topic vector $\boldsymbol{q}_g$ and last generated sentence embedding $\boldsymbol{r}_{i-1}$ for topic coherence as follows:

$$\boldsymbol{q}_g = \sum_{i=1}^{M} \alpha_i \boldsymbol{q}_i, \quad where\ \alpha_i = \frac{\|\boldsymbol{q}_i\|_2}{\sum_i \|\boldsymbol{q}_i\|_2}.$$

We first transform sentence embedding $\boldsymbol{r}_{i-1}$ to $\boldsymbol{q}_{i-1}^{h}$ through a 2-layer fully-connected unit. Then we merge $\boldsymbol{q}_{i-1}^{h}$ and $\boldsymbol{q}_i$ as follows:

$$\boldsymbol{q}_i^{'} = \beta_i \boldsymbol{q}_i + (1-\beta_i)\boldsymbol{q}_{i-1}^{h}, \beta_i = sigmoid(\boldsymbol{W}_{c,i}\boldsymbol{q}_i - \boldsymbol{W}_{c,i-1}\boldsymbol{q}_{i-1}^{h}).$$

The obtained vector $\boldsymbol{q}_i^{'}$ is then coupled with global topic vector $\boldsymbol{q}_g$ using a gating function implemented by a single-layer GRU.

$$\boldsymbol{q}_i^{c} = F_{GRU}^{c}(\boldsymbol{q}_i^{'}, \boldsymbol{q}_g).$$

Gate Module. Given the i-th topic vector $\boldsymbol{q}_i$, previous topic vector $\boldsymbol{q}_{i-1}$, the gate module generate a distribution $\boldsymbol{u}_i$ over $\{abnormal = 0, normal = 1\}$, that is:

$$\boldsymbol{u}_i = softmax(\boldsymbol{W}_u tanh(\boldsymbol{W}_{u,i}\boldsymbol{q}_i + \boldsymbol{W}_{u,i-1}\boldsymbol{q}_{i-1})).$$

Attentional Sentence Generator. Different from [2] that directly feed topic vector into generator, we enhance our sentence generator with a *topic guide attention module*. More specifically, coherent topic vector $\boldsymbol{q}_i^{c}$, regional visual feature $\boldsymbol{v}$ and previous hidden state $\boldsymbol{h}_{i,j-1}^{s}$ are fed into a fully connected layer, followed by a softmax to get the attention distribution over k regions as follows:

$$\mathscr{A} = \boldsymbol{W}_{att} tanh(\boldsymbol{W}_v^{att}\boldsymbol{v} + \boldsymbol{W}_t^{att}\boldsymbol{q}_i^{c}\boldsymbol{1}^k + \boldsymbol{W}_s^{att}\boldsymbol{h}_{i,j-1}^{s}\boldsymbol{1}^k).$$

$$\alpha_n = \frac{exp(\mathscr{A}_n)}{\sum_n exp(\mathscr{A}_n)}, \boldsymbol{c}_{i,j}^{s} = \sum_{n=1}^{k} \alpha_n \boldsymbol{v}_n.$$

The sentence generator is a single-layer LSTM which takes $\boldsymbol{e}_{i,j-1}, \boldsymbol{q}_i^{c}, \boldsymbol{c}_{i,j}^{s}$ as inputs:

$$\boldsymbol{h}_{i,j}^{s} = F_{RNN}^{s}(\boldsymbol{e}_{i,j-1}, [\boldsymbol{q}_i^{c}, \boldsymbol{c}_{i,j}^{s}], \boldsymbol{h}_{i,j-1}^{s}),$$

$$\boldsymbol{a}_{i,j} = softmax(\boldsymbol{W}_s\boldsymbol{h}_{i,j}^{s} + \boldsymbol{b}_s), s_{i,j} = argmax(\boldsymbol{a}_{i,j}), \boldsymbol{e}_{i,j} = \mathbf{E}\boldsymbol{1}_{s_{i,j}}.$$

where $\mathbf{E}$ is learnable embedding matrix and $\boldsymbol{1}_{y_{i,j}}$ is one-hot vector.

3 Experiments and Analysis

Settings. We evaluate the proposed A3FN method on public IU X-Ray dataset [1]. There are 2914 reports associated with 5828 images after filtering out reports without two complete image views. We randomly pick 10% reports as testing set. We compare the proposed A3FN with state-of-the-art report generation methods including CNN-RNN [3], Soft-ATT [4], ATT-RK [6], Co-ATT [2] and Multi-Modal [5]. For comparison, we reimplement baseline models [2,5] using Pytorch since the codes of these models are not available. We report the performance of all models on frequently used metrics BLEU-{1,2,3,4}, CIDEr and ROUGE in Table 1. We compute the keywords accuracy (KA) used in [5] and positive abnormality terminology detection accuracy (Acc) as a measurement of abnormal detection. Table 2 shows evaluation results of KA and Acc.

Table 1. Automatic evaluation results.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE
IU X-Ray	CNN-RNN [3]	0.311	0.218	0.137	0.092	0.124	0.262
	Soft-ATT [4]	0.351	0.237	0.161	0.12	0.278	0.314
	ATT-RK [6]	0.341	0.221	0.153	0.106	0.187	0.302
	Co-ATT [2]	0.421	0.324	0.225	0.174	0.331	0.341
	Multi-Modal [5]	0.434	0.331	0.234	0.177	0.312	0.346
	A3FN	**0.443**	**0.337**	**0.236**	**0.181**	**0.374**	**0.347**

Table 2. Positive abnormality terminology detection accuracy and KA

Dataset	Model	KA (%)	Acc (%)
IU X-Ray	Multi-Modal [5]	58.7	10.1
	Co-ATT [2]	57.3	11.37
	A3FN	**61.0**	**13.25**

Results and Analysis. The proposed A3FN method outperforms all other baselines. Compared with these hierarchical models [2,5], our A3FN method outperforms Co-ATT [2] and Multi-Modal [5] by a large margin, with respectively **11.3%** and **19.9%** relative improvement on CIDEr score. Furthermore, the proposed A3FN model achieves best KA score and best positive abnormality detection accuracy score, which means our generated reports detect more positive abnormality and cover more topics.

4 Conclusion

In this paper, we propose a novel model named Attention-based Abnormal-Aware Attention Network (A3FN) which aims to generate structured, detailed, topic coherent, and abnormal-aware radiology reports. Our model achieves the competitive results compared with all state-of-the-art models.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China Projects No. U1636207, No. 91546105, the Shanghai Science and Technology Development Fund No. 16JC1400801, No. 17511105502, No. 17511101702, Suzhou Science and Technology Bureau Technology Demonstration Project (SS201712, SS201812).

References

1. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. JAMIA **23**, 304–310 (2015)
2. Jing, B., Xie, P., Xing, E.P.: On the automatic generation of medical imaging reports. In: Proceedings of ACL (2018)

3. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of CVPR (2015)
4. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of ICML (2015)
5. Xue, Y., et al.: Multimodal recurrent model with attention for automated radiology report generation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 457–466. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_52
6. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of CVPR (2016)