AIMNET: ADAPTIVE IMAGE-TAG MERGING NETWORK FOR AUTOMATIC MEDICAL REPORT GENERATION

Jijun Shi[1], Shanshe Wang[2], Ronggang Wang[1], Siwei Ma[2]*

[1]School of Electronic and Computer Engineering, Peking University
[2]School of Computer Science, Peking University
shijijun@pku.edu.cn

ABSTRACT

In recent years, medical report generation has received increasing research interest with the goal of automatically generating long and coherent descriptive paragraphs that can describe in detail the observations of normal and abnormal regions in the input medical images. Unlike general image captioning tasks, medical report generation is more challenging for data-driven neural models. This is mainly due to severe visual and textual data biases. To address these problems, we propose an *Adaptive Image-Tag Merging Network (AIMNet)* that first predicts the tags of diseases from the input image, and then adaptively merges the visual information of the input image and disease information from the disease tags to learn the disease-oriented visual features that can better represent abnormal regions of the input image, and thus can be used to alleviate data bias problems. The experiments and analyses on the public MIMIC-CXR and IU-Xray datasets show that our proposed AIMNet achieves the state-of-the-art results under all metrics and significantly outperforms previous models on the MIMIC-CXR and IU-Xray datasets with relatively 15.1% and 6.5% margins in terms of BLEU-4 score.

Index Terms— Medical Report Generation, Data Bias, Attention Mechanism

1. INTRODUCTION

Medical images like radiology and pathology images are widely-used in disease diagnosis [1, 2, 3] and treatment. In practice, writing medical reports is very time-consuming and tedious for experienced radiologists, and prone to error for inexperienced radiologists [1, 2]. Therefore, an automated medical report generation system can reduce the workload of radiologists by assisting them in clinical decision-making, so this system is urgently needed [2, 4, 5, 6, 7].

Most existing medical report generation models follows the standard image caption approaches [8, 9, 10, 11] and employ the end to end architecture, e.g., a CNN based image encoder followed by a report decoder based on LSTM. Nevertheless, direct application of image captioning methods to medical images suffers from the following problems:1) **Visual data bias**: normal images dominate the dataset while abnormal images dominate the dataset [12]. In addition, for each abnormal image, the normal region dominates the image rather than the abnormal region. Besides, the abnormal regions only make up a small portion of the entire image; 2) **Textual data bias**: In a medical report, the radiologist tends to describe all items in an image, making the description of the normal regions dominate the entire report. In addition, many similar sentences were used to describe the same normal areas. Thus, even the most advanced models [7] tend to generate plausible general reports with no obvious abnormality narrative, as well as some errors and repetitions that fail to describe rare but important abnormalities.

To alleviate the above problems and thus describe the abnormalities effectively, we notice that the disease tags capture explicit information about the abnormalities and provide a comprehensive view of the abnormalities, while the visual information of the images contains the visual details of the abnormalities, such as location, severity and shape of the abnormalities. Therefore, we believe that the disease tags can be used as a guide to help the model describe the full range of abnormalities during report generation, while the visual information can be used to describe the more specific visual details of each abnormality. To this end, we propose a new Adaptive Image-Tag Merging Network (AIMNet). In its implementation, AIMNet first predicts disease tags from the input images. Then, an importance-based merging mechanism based on the adaptive merging gate, is introduced to adaptively determine the type of information should be rely on for generating a word, producing disease-oriented visual features which can represent abnormal regions of the input image and thus can be used to alleviate data bias problems. In summary, our approach can generate structured reports that are significantly consistent with ground truth reports and are supported by accurate abnormal descriptions.

In short, the contributions of this paper are as follows:

- In this paper, we propose the AIMNet ,which contains a importance-based merging mechanism, to alleviate the problem of data bias in order to generate accurate med-

*Corresponding author.

ical reports guided by disease tags.

- Experiments and analyses on two public benchmark datasets demonstrate the validity of our arguments and the effectiveness of our approach, which achieves state-of-the-art performance at all metrics.

2. RELATED WORKS

The generation of medical reports is similar to image captioning [13, 14], whose purpose is to generate a sentence to describe the image. In image captioning, encoder-decoder framework [8] has achieved great success [10, 15, 8, 16, 17]. Encoders [18] perform the computing of visual representation of images, then decoders [19, 20, 21] generate target sentences based on image representation. Given this success, most existing medical report generation models attempt to generate reports using the encoder-decoder framework in image captioning [2, 5]. However, the purpose of medical report generation is not to produce a single sentence, but to produce long paragraphs containing multiple sentences describing the normal and abnormal parts. In addition, due to data bias, these models tend to generate plausible but general reports without significant abnormal narratives [5, 22, 4].

3. APPROACH

We first formulate the problem; next, we introduce our proposed Adaptive Image-Tag Merging Network (AIMNet).

3.1. Problem Formulation

Given a medical image encoded as visual features V and disease Tags T, the goal of our framework is to generate a coherent report R describing detailed observations of normal and abnormal regions, which can be defined as:

$$\text{AIMNet} : \{V, T\} \rightarrow R.$$

In the implementation, for visual feature V, we follow [5, 4, 23] to extract 2,048 7×7 feature maps by using ResNET-50 [18] pre-trained on ImageNet [24] and fine-tuned on CheXpert dataset [25], which are further projected to 512 7×7 feature maps, denoted as $V = \{v_1, v_2, ..., v_{N_\text{v}}\} \in \mathbb{R}^{N_\text{v} \times d}$ ($N_\text{v} = 49, d = 512$). Furthermore, we follow [2] to further predict the disease Tags T of the input image. Specifically, we further feed the extracted visual features T into a multi-label classification network, which is pre-trained on the downstream dataset as a multi-label classification task to generate a distribution of all pre-defined disease tags. Finally, the embeddings of the $N_\text{T} = 5$ most likely disease tags $T = \{d_1, d_2, ..., d_{N_\text{T}}\} \in \mathbb{R}^{N_\text{T} \times d}$ are used as the disease tags for the current input image. Based on the extracted V and T, at each time step t, our framework first adopts existing visual attention [8] and tag attention [26, 2] to capture salient visual information $v_t \in \mathbb{R}^d$ and salient disease information $d_t \in \mathbb{R}^d$. Then, we propose an importance-based merging mechanism, i.e., an adaptive merging gate, which effectively merges the visual information in the image with the disease information in the disease tags to generate disease-oriented visual features $c_t \in \mathbb{R}^d$. Next, We introduce feed-forward network (FFN) in the original Transformer decoder [20] to generate the final medical report based on the c_t. Briefly, our AIMNet framework is formulated as:

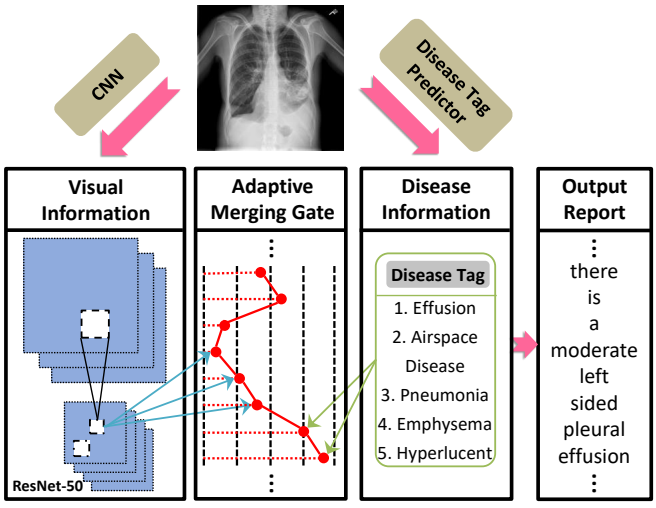

Fig. 1. Illustration of the AIMNet. The visual information captured by CNN and the disease information in the extracted disease tags are first condensed by attention mechanisms respectively. The adaptive merging gate then adaptively adjusts the weight between the visual information and the disease information for producing disease-oriented visual features to alleviate the data bias problem, resulting in generating accurate medical reports under the guidance of the disease tags.

$$\text{Visual Attention} : V \rightarrow v_t; \quad \text{Tag Attention} : T \rightarrow d_t;$$
$$\text{Adaptive Merging Gate} : \{v_t, d_t\} \rightarrow c_t; \quad \text{FFN} : c_t \rightarrow R.$$

Through the above process, our framework is able to alleviate the data bias problem and generate accurate and robust reports. During training, given ground truth reports of input images, we can train the model by minimizing supervised training losses such as cross-entropy loss.

3.2. Adaptive Image-Tag Merging Network (AIMNet)

The Fig. 1 illustrates our proposed method. In implementation, since medical report generation needs to generate a long paragraph, we use Transformer [20] as the basic module of our Adaptive Image-Tag Merging Network (AIMNet) and introduce Visual attention, Tag attention and adaptive merging gate to efficiently generate robust and accurate reports. In particular, the Transformer consists of a multi-head attention

Table 1. Results on the MIMIC-CXR and IU-Xray datasets. Higher value denotes better performance in all columns. As we can see, the AIMNet outperforms existing models substantially under all metrics, which proves the effectiveness of our approach.

Datasets	Methods	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MIMIC-CXR	CNN-RNN [8]	2015	29.9	18.4	12.1	8.4	12.4	26.3
	AdaAtt [15]	2017	29.9	18.5	12.4	8.8	11.8	26.6
	Up-Down [10]	2018	31.7	19.5	13.0	9.2	12.8	26.7
	R2Gen [7]	2020	35.3	21.8	14.5	10.3	14.2	27.7
	PPKED [23]	2021	36.0	22.4	14.9	10.6	14.9	28.4
	AIMNet	Ours	**38.8**	**24.1**	**16.0**	**12.2**	**16.7**	**29.3**
IU-Xray	HRGR-Agent [4]	2018	43.8	29.8	20.8	15.1	-	32.2
	CMAS-RL [5]	2019	46.4	30.1	21.0	15.4	-	36.2
	SentSAT+KG [6]	2020	44.1	29.1	20.3	14.7	-	36.7
	R2Gen [7]	2020	47.0	30.4	21.9	16.5	18.7	37.1
	PPKED [23]	2021	48.3	31.5	22.4	16.8	19.0	37.6
	AIMNet	Ours	**49.2**	**32.0**	**23.6**	**17.9**	**21.8**	**39.5**

(MHA) and a feed-forward network (FFN)[1].

At each time step t, to generate each word y_t in the final report, our model first takes as input the embedding of the current input word $x_t = w_t + e_t$ (w_t : word embedding, and e_t : fixed position embedding[20]) to obtain the current hidden state $h_t = \text{MHA}(x_t, x_{1:t}) \in \mathbb{R}^d$. Then, visual attention and tag attention are introduced to capture salient visual information $v_t \in \mathbb{R}^d$ and salient disease information $d_t \in \mathbb{R}^d$:

$$v_t = \text{Visual Attention}(h_t, V) = \text{MHA}(h_t, V); \quad (1)$$

$$d_t = \text{Tag Attention}(h_t, T) = \text{MHA}(h_t, T). \quad (2)$$

Adaptive Merging Gate When the decoder generates different types of words, it is not reasonable to treat visual information v_t and disease information d_t equally. For example, when generating abnormal words (such as effusion and scoliosis), d_t should be more important than v_t because it contains clear abnormal information that is responsible for a comprehensive view. However, when generating the location, severity, and shape of abnormalities, v_t is more important because it contains visual information describing the visual details of the abnormalities. Therefore, we introduce a new score-based merging mechanism to adaptively adjust the balance:

$$\gamma_t = \sigma(S(d_t) - S(v_t)) \quad (3)$$

$$c_t = \gamma_t d_t + (1 - \gamma_t) v_t, \quad (4)$$

where σ is the sigmoid function, $\gamma_t \in [0, 1]$ denotes the importance of d_t compared to v_t, and S is a scoring function to evaluate the importance of disease information and visual information. To achieve this, we empirically find that a multilayer perceptron (MLP), i.e., FC-ReLU-FC, can efficiently simulate the scoring function.

Lastly, the disease-oriented visual feature $c_t \in \mathbb{R}^d$ is used to predict the next word: $y_t \sim p_t = \text{Softmax}([c_t; h_t]W^p + b^p)$, where W^p and b^p are learnable parameters; [;] represents the concatenation operation; each value of $p_t \in \mathbb{R}^{|D|}$ is a probability indicating how likely the corresponding word in the vocabulary D is the current output word. Given that the ground truth report $R^* = \{y_1^*, y_2^*, ..., y_{N_R}^*\}$ provided by the radiologist, we can minimize the widely used cross-entropy loss : $L_{\text{XE}}(\theta) = -\sum_{i=1}^{N_R} \log(p_\theta(y_i^* | y_{1:i-1}^*))$. In summary, the model is encouraged to alleviate the data bias problem and thereby generating accurate medical reports.

[1]Due to limited space, please refer to [20] for detailed introduction.

4. EXPERIMENTS

In this section, we first introduce the benchmark datasets and metrics used as the basis for our experiments, as well as the experimental settings that we tested. We subsequently show evaluations of our proposed approach.

4.1. Datasets, Metrics and Settings

We conduct experiments on two public datasets, namely MIMICCXR [27] and IU-Xray [28]. MIMIC-CXR included 377,110 chest X-ray images and 227,835 reports from 64,588 patients. Following [7], we report our results using the official split, yielding 368,960 in the training set, 2,991 in the validation set, and 5,159 in the test set. IU-Xray contains 7,470 chest X-ray images associated with 3,955 reports. Following previous work [7, 5, 4], we randomly split the dataset into 70%-10%-20% train-validation-test splits. For a fair comparison with existing models [4, 7], we employ the standard evaluation toolkit [13] to compute widely used metrics, namely BLEU, METEOR and ROUGE-L, measure how well the generated reports match the ground truth reports. Since our method is based on Transformer [20], we follow the original setting in [20] and set $d = 512$. we use the Adam optimiser with a batch size of 16 and a learning rate of 2e-4 at a maximum of 100 epochs of parameter optimization.

Table 2. Quantitative analysis of our proposed method, which includes the Visual Attention (VA), the Tag Attention (TA) and the Adaptive Merging Gate (AMG). The Base Model and Full Model denote the ResNet-50 image encoder [18] equipped with the Transformer report decoder [20] and our proposed AIMNet (i.e., 'w/ VA+TA+AMG'), respectively.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Base Model	44.2	28.5	19.5	13.8	17.6	35.7
w/ VA	46.4	29.7	20.8	15.5	18.1	36.6
w/ TA	45.3	29.4	20.3	14.6	17.9	36.2
w/ VA + TA	47.2	30.3	21.4	16.1	19.0	36.9
Full Model	**49.2**	**32.0**	**23.6**	**17.9**	**21.8**	**39.5**

4.2. Main Results

Several representative models are selected for comparison, including three recently released state-of-the-art models, namely PPKED [23], R2Gen [7], and SentSAT+KG [6]. The results on the two datasets are shown in Table 1. As we can see, our AIMNet achieves the best results on both datasets under all metrics and outperforms previous state-of-the-art models on MIMIC-CXR and IU-Xray datasets and has relative 15.1% and 6.5% margins in BLEU-4 scores. It validates the effectiveness of our proposed AIMNet in generating accurate and appropriate reports, which can help radiologists make decisions and reduce their workload.

4.3. Quantitative Analysis

We investigate the contribution of each component to the IU-Xray dataset in our approach. As can be seen from Table 2, the introduced visual attention and tag attention improve performance on all metrics, demonstrating the effectiveness of our method in capturing salient visual information and salient disease information.For the full model further incorporating the proposed importance-based merging mechanism, i.e., Adaptive Merging Gate (AMG), Table 2 shows that AMG can significantly improve the performance under all metrics, with a BLEU improvement of up to 11.2%-4 points. This demonstrates the effectiveness of our AMG in adaptively determining which information to rely on when generating words to produce disease-oriented visual features, which is critical for alleviating data bias problems and improving the performance of medical report generation.

4.4. Qualitative Analysis

Visualization Fig. 1 shows a visualization of our AIMNet. As we can see, disease information is active when generating phrases (e.g. *pleural effusion*) that contain relevant disease tags (e.g. *effusion*). The merging gate learn to steer the flow of information efficiently. When generating words such as medium, left, and side, it prefers visual information, which contains the visual details of the abnormality, i.e., the location, severity, and shape of the abnormalities.

Error Analysis We find that errors mainly occurred when there were incorrectly predicted disease tags. AIMNet mistake erroneous disease tags for appropriate tags during final report generation. More robust disease tag predictors may help, but are unlikely to be avoided entirely [2].

5. CONCLUSION

In our work, we propose a novel AIMNet to address the problem of data bias in medical report generation by generating disease-oriented visual features. Experiments show that our approach not only generates meaningful reports with accurate abnormal descriptions, but also achieves state-of-the-art results at all metrics and outperforms the previous models on the MIMIC-CXR and IU-Xray datasets with a relative rate of 15.1% and 6.5% margins in terms of BLEU-4. The analysis further validates the effectiveness of our approach in addressing data bias to describe rare and significant abnormalities.

6. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (62072008, 62025101) and High-performance Computing Platform of Peking University, which are gratefully acknowledged. We also sincerely thank all the anonymous reviewers and chairs for their constructive comments and suggestions that substantially improved this paper. Siwei Ma is the corresponding authors of this paper.

7. REFERENCES

[1] A. Brady, R. Ó. Laoide, Peter Mccarthy, and R. Mcdermott, "Discrepancy and error in radiology: Concepts, causes and consequences," *The Ulster Medical Journal*, vol. 81, pp. 3 – 9, 2012.

[2] Baoyu Jing, Pengtao Xie, and Eric P. Xing, "On the automatic generation of medical imaging reports," in *ACL*, 2018.

[3] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun, "Contrastive attention for automatic chest x-ray report generation," in *ACL/IJCNLP (Findings)*, 2021.

[4] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *NeurIPS*, 2018.

[5] Baoyu Jing, Zeya Wang, and Eric P. Xing, "Show, describe and conclude: On exploiting the structure information of chest x-ray reports," in *ACL*, 2019.

[6] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu, "When radiology report generation meets knowledge graph," in *AAAI*, 2020.

[7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan, "Generating radiology reports via memory-driven transformer," in *EMNLP*, 2020.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[9] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun, "simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions," in *EMNLP*, 2018.

[10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and VQA," in *CVPR*, 2018.

[11] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Kai Lei, and Xu Sun, "Exploring and distilling cross-modal information for image captioning," in *IJCAI*, 2019.

[12] Hoo-Chang Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *CVPR*, 2016.

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[14] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun, "Prophet attention: Predicting attention with future attention," in *NeurIPS*, 2020.

[15] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.

[16] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou, "Federated learning for vision-and-language grounding problems," in *AAAI*, 2020.

[17] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *NeurIPS*, 2019.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[21] Fenglin Liu, Xuancheng Ren, Guangxiang Zhao, and Xu Sun, "Layer-wise cross-view decoding for sequence-to-sequence learning," *arXiv preprint arXiv:2005.08081*, 2020.

[22] Fenglin Liu, Shen Ge, and Xian Wu, "Competence-based multimodal curriculum learning for medical report generation," in *ACL/IJCNLP*, 2021.

[23] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *CVPR*, 2021.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI*, 2019.

[26] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[27] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng, "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[28] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Medical Informatics Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.