

Attention Enhanced Network with Semantic Inspector for Medical Report Generation

Anonymous

Anonymous Organization **@*****.***

Abstract. Medical report generation can be helpful in diagnoses. Despite the previous efforts of researchers, current models still need improvements in extraction of image features and quality of generated reports. In this paper, we propose an attention enhanced network with semantic inspector (AENSI) as a new automatic medical report generation model, which serves to help doctors to get a high-quality report. For the model, we adopt double-weighted multi-head attention as our attention module, where different heads are aggregated with double weights (DW-MHA) to enhance its power in catching subtle features and drawing correlations between images and texts. To prevent the drawback of imprecise multi-label classification modules used in current generation models, we design a novel module following decoder that treats tags as inspectors of the generated reports, namely Tag Inspector, as a substitute for previous classification module. Experimental results of AENSI achieve to the level of state-of-the-art. On IU X-ray, our model surpasses all previous works on every metrics; on PEIR Gross, our model ranks first on BLEU-4 and ROUGE and closely approaches the best on other metrics.

Keywords: medical report generation · image caption · Transformer · double-weighted multi-head attention · tag inspector

1 Introduction

Medical image reports are important in many disease diagnoses, but writing them by hands is time-consuming for doctors. A proposed way to release doctors from this burden is to introduce an accurate automatic medical image reports generation model. Medical reports generation derives from a more general task, image caption, whose models are mainly consist of a CNN encoder and a RNN decoder[18,20,8,1,5]. In 2017, Transformer [16] abandoned this framework and brought up a novel structure based on pure self-attention module, which performed well in plenty of works. D’Ascoli took advantage of both CNN and Transformer and proposed ConViT for better image features catch[4]. Researchers introduced these great works into medical report generation task, and made some specific improvements[7,22,24,12,21,6]. To strengthen the model’s competence in extracting visual features and generating longer descriptions, Tang et al. [15] designed IFE module and hierarchical decoder (H-Decoder).

Though these works made great progress in medical report generation, some challenges remain unsolved. First, attention mechanism enables the model to follow specific parts of the input image, but lacks information about which aspect the model should attach more emphasis. Second, with the complexity of multi-label classification with huge amount of candidate labels, it is difficult to obtain a precise classification results, leading to absence or negative influence in following reports generation.

Therefore, we raise a new model based on the standard encoder-decoder structure, AENSI, and propose two original innovations for the above problems. For the first problem, we use the multi-head attention structure with double weights (DW-MHA). Compared to traditional attention module, multi-head attention calculates the attention weights from various aspects and provides richer and more precise visual and linguistic features. We adopt double weights for head aggregation to inform the model of the important discrepancies between different heads. For the second problem, we propose Tag Inspector to utilize tags information effectively and reduce the model sensitiveness toward tags when generating reports. Fig. 1 illustrates the whole structure of our model.

Our contributions can be summarized as follow:

- (1) We enhance multi-head attention with double weights for better image-text correlation.
- (2) We design a novel module, tag inspector, to totally prevent the drawbacks of imprecise label classification for the first time.
- (3) We conduct thorough experiments on two widely-used datasets, and our model approaches and even exceeds the state-of-the-art works.

2 Methods

Overview Our model is constructed with encoder-decoder framework. Most encoders are based on the classic convolution architecture. Its “hard” inductive biases enable it to perform well on relatively small datasets, which comes at a cost of a potentially lower performance ceiling. Recently, Transformer received competitive results in CV (ViT [3]), and even surpassed convolution-based structures in some tasks. Nevertheless, its insensitivity to position and strong reliance on large training datasets are conspicuous defects. Stephane d’Ascoli et al. combined them in a convolutional-like ViT architecture called ConViT [4]. ConViT achieves sample-efficient and performance-improved learning simultaneously, which highly meets the requirement of medical report generation task with small samples yet strict demands of features extraction.

To enforce the visual features extraction ability of the model, SVEH-Net[15] applies an image feature encoding (IFE) module which gains global information while reserving local information simultaneously. We make some modulation to adapt our model. IFE accepts F_{CV_i} as input, and sends it into a fully connected layer to extract valuable correlations between different patches of the whole image. The output of this fully connected layer is the final visual feature F_O .

$$F_{CV_i} = \text{ConViT}(I) \quad (1)$$

$$\mathbf{F}_O = \text{IFE}(\mathbf{F}_{Cvi}) \quad (2)$$

where I is the input image, \mathbf{F}_{Cvi} is the extracted visual features.

H-Decoder[15] performs well in extract more semantic features and generate accurate reports. It contains two LSTMs, the first one targets encoding tag features, and the second one aims at generating sentences. Since the second LSTM decodes twice, it will generate two paragraphs, making up the final report jointly. This work performs well in generating medical reports, so we build our decoder based on H-Decoder.

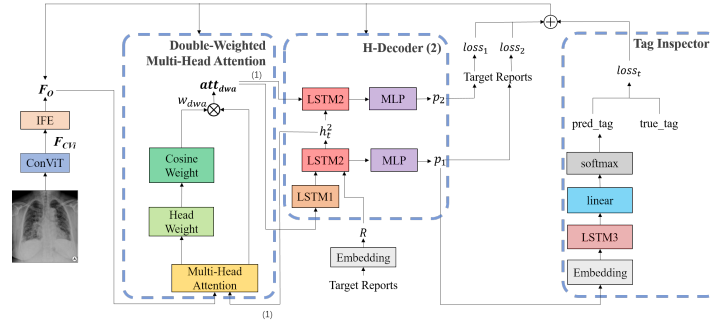


Fig. 1. The whole architecture of AENSI. (1) H-Decoder invokes the attention mechanism twice independently during its work.

Double-Weighted Multi-Head Attention (DW-MHA) Self-attention proposed in Transformer[16] calculates the relationship between two vectors after mapping them through three matrices Q , K , and V . Based on self-attention, which contains merely one head, multi-head attention [16] splits one Q , K and V into N parts ($N > 1$), each of which focuses on different aspects to enrich model information. We adopt it as the first step of our attention module.

$$\mathbf{att}_a^i = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

$$\mathbf{att}_a = \text{Concat}(\mathbf{att}_a^1, \mathbf{att}_a^2, \dots, \mathbf{att}_a^N)W \quad (4)$$

where N is the number of heads, d_k is the scaling factor, and \mathbf{att}_a is the output of multi-head attention.

Traditional multi-attention aggregates N heads by direct concatenation or addition, treating every head equally in the final reports generation. However, when we human observe a picture for a specific intention, we pay more attention to those highly correlative aspects and less attention to others. Inspired by this intuition, we propose weighted multi-head attention. Specifically, \mathbf{att}_a will be

mapped linearly by a group of weights \mathbf{w}_a . At the beginning of each training iteration, $\mathbf{w}_{a(i)}$ (i represents the training iteration) is calculated by $\mathbf{w}_{a(i-1)}$ with softmax function as given in Eq. (5). For the first start of model training, we assume each head shares equal importance, so $\mathbf{w}_{a(0)}$ is initialized to a straight 1 vector. To keep its consistency throughout the whole training process, $\mathbf{w}_{a(i)}$ will multiply a normalization factor N .

$$\mathbf{w}_{a(i)} = \text{softmax}(\mathbf{w}_{a(i-1)}) * N \quad (5)$$

$$\mathbf{att}_{\mathbf{w}_a(i)} = \mathbf{w}_{a(i)} * \mathbf{att}_{a(i)} \quad (6)$$

where $\mathbf{att}_{\mathbf{w}_a}$ is the output of attention module with single weights (W-MHA). The experiment on attention head weights (Table 1) confirms the difference

Table 1. Comparison of single weight and double weight on the IU X-ray. The highest value of each row is highlighted in boldface.

:single weight
 :double weight

Fold	Head 1	Head 2	Head 3	Head 4	Head 5	Head 6	Head 7	Head 8
Fold 1	2.737×10^{-3}	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	5.417×10^{-3}	14.964	5.444×10^{-3}	5.419×10^{-3}	5.508×10^{-3}	5.501×10^{-3}	5.407×10^{-3}	5.425×10^{-3}
Fold 2	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	14.964	5.392×10^{-3}	5.476×10^{-3}	5.599×10^{-3}	5.418×10^{-3}	5.490×10^{-3}	5.417×10^{-3}	5.458×10^{-3}
Fold 3	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}
	5.511×10^{-3}	5.473×10^{-3}	5.452×10^{-3}	5.425×10^{-3}	5.523×10^{-3}	5.444×10^{-3}	14.964	5.512×10^{-3}
Fold 4	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}	2.737×10^{-3}
	5.425×10^{-3}	5.376×10^{-3}	5.455×10^{-3}	5.449×10^{-3}	5.392×10^{-3}	14.964	5.388×10^{-3}	5.338×10^{-3}
Fold 5	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	14.964	5.436×10^{-3}	5.573×10^{-3}	5.399×10^{-3}	5.415×10^{-3}	5.466×10^{-3}	5.484×10^{-3}	5.506×10^{-3}
Fold 6	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981
	5.526×10^{-3}	5.575×10^{-3}	5.525×10^{-3}	5.506×10^{-3}	5.570×10^{-3}	5.486×10^{-3}	5.558×10^{-3}	14.964
Fold 7	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}
	5.567×10^{-3}	5.519×10^{-3}	5.489×10^{-3}	5.431×10^{-3}	5.531×10^{-3}	5.524×10^{-3}	14.964	5.486×10^{-3}
Fold 8	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	5.464×10^{-3}	5.474×10^{-3}	5.495×10^{-3}	5.514×10^{-3}	14.964	5.542×10^{-3}	5.480×10^{-3}	5.495×10^{-3}
Fold 9	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	5.422×10^{-3}	5.662×10^{-3}	5.387×10^{-3}	14.964	5.342×10^{-3}	5.435×10^{-3}	5.445×10^{-3}	5.390×10^{-3}
Fold 10	2.737×10^{-3}	2.737×10^{-3}	7.981	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}	2.737×10^{-3}
	5.591×10^{-3}	5.502×10^{-3}	14.964	5.530×10^{-3}	5.523×10^{-3}	5.517×10^{-3}	5.506×10^{-3}	5.592×10^{-3}

among heads. Particularly, several heads that are more relevant to the report generation attain conspicuously high weights after training, whereas other heads play comparatively subtle roles. To strengthen the advantages of those salient heads, we introduce the second weight. At each step, the head with the largest \mathbf{w}_a (i.e. the most important head) was chosen as base. Cosine weight ($\cos(i)^j$) represents the cosine similarity between head j and base at i iteration. For each batch, the cosine similarity of the same head will sum up and divide by the number of heads to compute the second weight $w_{\cos(i)}^j$.

$$\cos(i)^j = \cos(\mathbf{att}_{a(i-1)}^j, \mathbf{att}_{a(i-1)}^{\text{base}}) \quad (7)$$

$$w_{cos(i)}^j = \frac{\sum_{k \in i} cos(i)_k^j}{N} \quad (8)$$

$$w_{dwa(i)}^j = w_{a(i)}^j * (2 - w_{cos(i)}^j) \quad (9)$$

$$\mathbf{att}_{dwa(i)}^j = w_{dwa(i)}^j * \mathbf{att}_a^j(i) \quad (10)$$

where $cos(i)^j \in [-1, 1]$, $cos(i)_k^j$ is the cosine similarity between head j and *base* of data k , N is the number of heads, \mathbf{att}_{dwa} is the output of attention module after double weighted (the final attention output in our model). Apparently, the weight of *base* is free from cosine weight. The final attention \mathbf{att}_{dwa} is calculated by both weights.

In our experiment, the batch size and the number of heads are set to 16 and 8 separately, so the final head weights close to 2 times of single weights statistically with higher flexibility compared to directly double them. The second weight further emphasizes the most important heads and leads to better results in report generation.

Tag Inspector Medical image datasets often gives tags indicating disease type and other valuable information explicitly. Current models employ tags through a multi-label classification structure for tag prediction, and generate reports with its outputs. This is reasonable if tag prediction is precise, but due to the small amount of most medical datasets with a large number of tags, it is difficult to train an effective image classification model. Since reports are generated directly based on tags, inaccurate predictions will be useless or even act virtually in model performance.

Tags can be seen as key words of reports. Inspired by this relationship, we propose a novel method to utilize tags effectively. We adopt tags as “quality inspector” of generated reports during the training process rather than engaging them directly in report generation. It also consists of a multi-label classification module; however, different from previous works, its input is the generated report rather than the image. This module contains a tag prediction LSTM and a classification layer. Embedded reports R_{emb} are sent into the tag prediction LSTM, and it outputs predict tags $ptag$. Then, we flatten $ptag$ with a linear mapping layer and gain probability of each tag with softmax.

$$ptag' = \text{LSTM}(R_{emb}) \quad (11)$$

$$ptag = \text{softmax}(\mathbf{W}_t * ptag' + b_t) \quad (12)$$

where \mathbf{W}_t and b_t are trainable parameters used to flatten $ptag'$. This “inspector” participates in the training process by adding its loss $loss_t$ in backward propagation.

$$\begin{aligned} loss_t = & -\frac{1}{C} * \sum_i tag_i * \log((1 + \exp(-ptag_i))^{-1}) \\ & + (1 - tag_i) * \log\left(\frac{\exp(-ptag_i)}{(1 + \exp(-ptag_i))}\right) \end{aligned} \quad (13)$$

$$loss = loss_1 + \lambda * loss_2 + 5 * loss_t \quad (14)$$

where tag represents the true tags, $i \in \{0, \dots, n-1\}$, $tag_i \in \{0, 1\}$; n is the number of tags types. $loss$ is the final loss of model, $loss_1$ ($loss_2$) is the cross-entropy loss between reports generated by first (second) LSTM2 in H-Decoder and true captions. The second LSTM2 generation is confined by an adjustable parameter $\lambda \in (0, 1]$. Since $loss_t$ is merely one-tenth of $loss_1$ and $loss_2$ statistically, $loss_t$ needs to be amplified to balance. Given that $loss_t$ does not directly reflect to reports quality like $loss_1$ and $loss_2$, we set its coefficient at 5.

3 Experiments

3.1 Datasets, Metrics and Implementation Details

Datasets and Metrics We test out model on two widely-used datasets. **IU X-ray** is a chest X-ray collection selected from the Indiana Network for Patient Care by researchers from Indiana University (<https://openi.nlm.nih.gov/>, [2]). It contains 7,470 radiology images and 3,995 reports. Reports consist of five parts: *comparison*, *indication*, *findings*, *tags*, and *impression*. We adopt 6,730 images and concatenate *Findings* and *impression* as target captions following the preprocessing method in [14].

To further test our model’s ability on identifying diseases on other organs and generating short reports, we conduct experiments on **PEIR Gross**, the Gross sub-collection of the Pathology Education Informational Resource (PEIR) digital library (<https://peir.path.uab.edu/library/>). It is a public medical image library for medical education. We obtain 7,442 image-caption pairs from the Gross sub-collection and preprocess it following [15]. We adopt 10-fold cross-validation on both datasets, where each fold contains 500 randomly selected non-overlapped images. The rest images are always treated as training set.

We choose three widely-used metrics to evaluate our work: BLEU [13], ROUGE-L [11], and CIDEr [17]. In the tables, we notate n-gram BLEU as B-n, while ROUGE-L and CIDEr are simplified to R and C.

Implementation Details We apply ConViT pretrained on ImageNet without the last classification layer to extract 512-dimension visual features. The dimension of word embedding and the dimension of the hidden state of all LSTM are set to 512. We adopt ADAM [9] as the optimizer throughout our model. The learning rate is 0.0004. According to the exploration in [15], we set λ to 0.5 to reach its best performance. During the evaluation, we use the beam search strategy. Metrics computation is implemented with the widely used image captioning tool¹.

¹ <https://github.com/Maluuba/nlg-eval>

Table 2. Comparison of Proposed Methods with state-of-the-art Methods on the IU X-ray and PEIR Gross. The highest value of each column is highlighted in bold and the second highest is highlighted in blue. ”-” means the corresponding metrics are not provided in the original article. ($\times 100$)

Dataset	Model	B-1	B-2	B-3	B-4	R	C
IU X-ray	co-att[7]	46.2	33.1	24.2	17.8	40.5	40.8
	nearest-neighbor[14]	28.1	15.2	09.1	05.7	20.9	-
	KERP[10]	48.2	32.5	22.6	16.2	33.9	28.0
	MvH[23]	43.6	31.2	22.9	17.0	37.2	32.8
	A3FN[19]	44.3	33.7	23.6	18.1	34.7	-
	TriNet[21]	47.8	34.4	24.8	18.0	39.8	43.9
	TransGen[6]	46.1	28.5	19.6	14.5	36.7	-
	PPKED[12]	48.3	31.5	22.4	16.8	37.6	35.1
	SVEH-Net[15]	50.8	35.6	25.9	19.1	40.8	41.5
	AENSI (ours)	54.2	36.4	26.7	19.8	43.3	46.4
Dataset	Model	B-1	B-2	B-3	B-4	R	C
PEIR Gross	co-att[7]	30.0	21.8	16.5	11.3	27.9	32.9
	nearest-neighbor[14]	34.6	26.2	20.6	15.6	34.7	-
	SVEH-Net[15]	46.6	32.3	23.3	16.9	37.4	26.9
	AENSI (ours)	44.2	31.5	22.6	17.4	43.5	28.2

3.2 Results and Analization

Comparison to State-of-the-Art We compare our model with well-performed previous works (Table 2). Our model obtains state-of-the-art results on both validation datasets. In particular, on IU X-ray, our model reaches the highest in all metrics; on PEIR Gross, our model reaches the highest in BLEU-4 and ROUGE-L, and almost achieves the best in the rest. This proves that our model is more powerful in long paragraphs generation, but also competitive in short sentences.

Table 3. Ablation study of key structures on IU X-ray. Baseline model consists of a ViT encoder, H-Decoder and traditional multi-head attention module. The highest value of each column is highlighted in bold. ($\times 100$)

Model	B-1	B-2	B-3	B-4	R	C
Baseline	49.5 \pm 1.7	33.2 \pm 1.3	23.6 \pm 1.3	17.2 \pm 1.3	39.8 \pm 1.8	40.6 \pm 4.8
+W-MHA	49.7 \pm 1.7	33.6 \pm 1.2	24 \pm 1.4	17.3 \pm 1.0	39.8 \pm 1.9	41.1 \pm 4.4
+DW-MHA	49.9 \pm 2.1	34.1 \pm 2.5	24.4 \pm 2.4	17.8 \pm 2.2	40.6 \pm 1.9	41.9 \pm 3.8
+Tag Inspector	49.8 \pm 2.8	33.8 \pm 2.5	24.3 \pm 2.1	17.9 \pm 1.5	40.4 \pm 1.9	41.8 \pm 2.5
+All (AENSI)	50.7\pm3.5	34.8\pm2.7	25.1\pm2.9	18.4\pm2.2	41.0\pm2.4	42.0\pm4.5

Ablation Study We also perform ablation experiments on IU X-ray to test the contribution of each proposed component (Table 3). From the comparison, we can observe a tendency for improvement with the addition of proposed modules. Finally, our complete model AENSI obtains the best result. Furthermore, the data between DW-MHA and Tag Inspector implies that their validities are close to each other. Besides, the gradual increase in the top three rows proves the effects of both heads in DW-MHA. The higher improvement amplitude after adding the second weights than merely adding single weights shows that two heads performing jointly is better than single head.



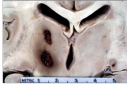
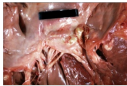
Dataset	Image & Tags	Ground Truth	Our Model
IU X-ray	 normal	No acute cardiopulmonary abnormality. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation.	No acute cardiopulmonary abnormality. The lungs are clear. There is no focal airspace opacity. No pleural effusion or pneumothorax. The cardiomeastinal silhouette is normal.
	 spine, degenerative	No acute cardiopulmonary process. Normal heart size and mediastinal contours. Lungs are clear. There is no pneumothorax or pleural effusion. Degenerative changes are seen in the spine.	No active disease. Both lungs are clear and expanded. Heart and mediastinum within normal limits. Degenerative changes in the thoracic spine.
PEIR Gross	 brain, gross, hemorrhage, nervous	Gross: nervous: brain: hemorrhage: gross fixed tissue close-up view of basal ganglia showing two hemorrhages.	Gross: nervous: brain: hemorrhage: gross fixed tissue coronal section cerebral hemispheres with typical hypertensive hemorrhage in basal ganglia.
	 cardiovascular, gross, heart, mitral, stenosis	Gross: cardiovascular: heart: mitral stenosis: gross natural color close-up view of typical rheumatic mitral stenosis.	Gross: cardiovascular: heart: joints: gross natural color close-up view of <unk> with <unk>.

Fig. 2. Examples of medical reports generated by AENSI on IU X-ray and Peir Gross.

4 Conclusion

In this paper, we provide two possible solutions to the major challenges of current medical report generation models. First, we adopt double-weighted multi-head attention to let the model focus on more meaningful parts of images in the generation process. Second, we propose Tag Inspector to make full use of tags information under the limitation of training samples. The wide and thorough experiments on both radiology and pathology datasets show the validity of our methods. We compare our model with current state-of-the-art works, which shows that our model reaches and even surpass their performance. In view of these success, the next step we plan to transplant our model to larger datasets such as MIMIC-CXR.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**(2), 304–310 (2016)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
4. d’Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning. pp. 2286–2296. PMLR (2021)
5. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4633–4642 (2019). <https://doi.org/10.1109/ICCV.2019.00473>
6. Jia, X., Xiong, Y., Zhang, J., Zhang, Y., Suzanne, B., Zhu, Y., Tang, C.: Radiology report generation for rare diseases via few-shot transformer. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1347–1352 (2021). <https://doi.org/10.1109/BIBM52615.2021.9669825>
7. Jing, B., Xie, P., Xing, E.P.: On the automatic generation of medical imaging reports. In: ACL (1) (2018)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015). <https://doi.org/10.1109/CVPR.2015.7298932>
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)
10. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6666–6673 (2019)
11. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 605–612 (2004)
12. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13753–13762 (2021)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
14. Pavlopoulos, J., Kougia, V., Androutsopoulos, I.: A survey on biomedical image captioning. In: Proceedings of the Second Workshop on Shortcomings in Vision and Language. pp. 26–36 (2019)
15. Tang, Q., Yu, Y., Feng, X., Peng, C.: Semantic and visual enrichment hierarchical network for medical image report generation. In: 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML). pp. 738–743. IEEE (2022)

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
17. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4566–4575 (2015)
18. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3156–3164 (2015)
19. Xie, X., Xiong, Y., Yu, P.S., Li, K., Zhang, S., Zhu, Y.: Attention-based abnormal-aware fusion network for radiology report generation. In: *International Conference on Database Systems for Advanced Applications*. pp. 448–452. Springer (2019)
20. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: *International conference on machine learning*. pp. 2048–2057 (2015)
21. Yang, Y., Yu, J., Zhang, J., Han, W., Jiang, H., Huang, Q.: Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Transactions on Multimedia* (2021)
22. Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., Zheng, Q.: Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: *2019 IEEE International Conference on Data Mining (ICDM)*. pp. 728–737. IEEE (2019)
23. Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 721–729. Springer (2019)
24. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, pp. 12910–12917 (2020)