Semantic and Visual Attention-Driven Multi-LSTM Network for Automated Clinical Report Generation

Cheng Huang[1], Junhao Shen[1]*, Beichen Hu[1]*, Mohammad Ausaf Ali Haqqani[1], Jia Zhang[1]

[1]Southern Methodist University
3101 Dyer Street
Dallas, Texas 75205 USA
{chenghuang, junhaos, beichenh, malihaqqani, jiazhang}@smu.edu

Abstract

Medical image processing has gained significant momentum in recent years. Latest advancements in machine learning and deep learning has enabled AI-powered generation of medical image reports. Some limitations remain, however, for example, generated reports may be lengthy without highlighting anomaly as desired, and some minor features might be neglected which fails in fine-grained labeling. To tackle the aforementioned challenges, this paper presents Semantic and Visual Attention-Driven Multi-LSTM Network (SVAML), a novel framework tailored to enhance medical image report generation. Specially, SVAML introduces a Double-Weighted Multi-Head Attention mechanism with a new weight function, to learn patterns of how to focus on describing important impressions from medical images. In addition, SVAML devises a Label Discriminator (LD), a module to learn intricate features to support more sensitive multi-label classification. Extensive experiments over two known public datasets, the IU X-ray dataset and the PEIR Gross dataset, have demonstrated the effectiveness of the presented SVAML framework.

Introduction

Computer-aided medicine and health care, such as medical reports generation and online medical pre-diagnosis, has been obtaining significant momentum in recent years. Medical reports generation typically comprises two categories: medical conversation-based report generation and medical image report generation. Medical image report generation refers to cross-disciplinary efforts at the intersection of computer vision, bioinformatics, pattern recognition, machine learning, and natural language processing (NLP). Latest advancements in deep learning has enabled AI-powered generation of medical image reports, leading to radiologist-level report generation.

Traditionally, professional radiologists read and interpret medical images, such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, or pathological imaging, and depict a diagnostic report typically comprising indications, findings and impressions. As medical images become one major instrument supporting disease diagnosis, and individual medical images may exhibit unforeseen abnormal features, manual analysis and report composition becomes not only time consuming but error prone. Automatic medical image report generation thus becomes highly demanded (Anderson et al. 2018).

Medical image report generation is rooted in image captioning (Farhadi et al. 2010) aiming to generate textual descriptions from images. Deep learning-powered works are usually centered around a CNN-RNN model, where a CNN learns visual patterns and generates a representation from an image, followed by an RNN generating textual descriptions from the intermediate representation (Anderson et al. 2018; Zhou, Li, and Liang 2020; Maitre, Bouchard, and Gaboury 2020).

Vaswani et al. introduce the attention-based Transformer architecture (Vaswani et al. 2017). Their attention mechanism allows neural networks to automatically learn and selectively focus on important information in the input, improving the performance and generalization ability of NLP models (Vaswani et al. 2017). Transformer is later applied to the computer vision field, to enhance detection (Liang et al. 2020) and instance segmentation (Gavrilyuk et al. 2020). In the realm of medical imaging processing, some researchers employ the Transformer to improve graph neural networks for processing pathology images to predict disease grade (Zheng et al. 2022b). Meanwhile, some researchers redesign a CNN-LSTM model based on Transformer, to extract domain data more effectively, thereby enabling human body data analysis and health monitoring (Chen et al. 2020). Furthermore, a pure transformer-based framework is designed to enhance the descriptive labels generation from medical images (Wang et al. 2022), integrating multi-label diagnostic classification and word importance weighting. Furthermore, Vision Transformer (ViT) (Dosovitskiy et al. 2021) splits an image into fixed-size patches, and feeds the linear sequence of the embeddings of each patch to a Transformer encoder. ConViT (d'Ascoli et al. 2021) further combines CNN and ViT by allowing each self-attention layer to decide whether to behave as a convolutional layer or not.

Despite of the substantial advancements achieved in medical imaging report generation, significant challenges remain and result in the omission of crucial details or even adversely affect the integrity of the generated reports (Vaswani et al. 2017). Among them, two limitations deserve investigation. Firstly, generated reports using existing models may

*These authors contributed equally.

be lengthy without highlighting anomaly as desired. Although the attention mechanism may grant a model capacity to concentrate on specific segments of an input image, it does not inherently discern the critical aspects, i.e., typically anomaly, to which it should ascribe greater significance (Gavrilyuk et al. 2020). Secondly, some minor features might be neglected during model learning, which leads to failure in fine-grained labeling. The inherent complexity of multi-label classification, aggravated by a vast array of potential labels, often impedes the attainment of precise classification outcomes (Liang et al. 2020).

In order to tackle the aforementioned two challenges, this paper introduces a novel framework called Semantic and Visual Attention-Driven Multi-LSTM Network (SVAML), tailored to enhance medical image report generation. Synergistically leveraging ConViT (d'Ascoli et al. 2021) and Muti-LSTM (Hochreiter and Schmidhuber 1997; Zheng et al. 2022a), SVAML introduces two new modules. The first module incorporates a multi-head attention structure with double weights. As opposed to the conventional attention module, this module improves the correlation between target features of medical images and the descriptions of corresponding labels. The integration of double weights for head aggregation enlightens the model about the pivotal variances amongst different attention heads. The second module is 'Label Discriminator' devised to harness tag information efficaciously and attenuate the model's over-sensitivity to tags during report generation. This module will mitigate the limitations associated with imprecise keyword identification and classification to achieve more effective information extraction.

The contributions of this paper can be summarized in three-fold:

- This paper augments the multi-head attention mechanism with a double weighting instrument, which learns patterns of how to focus on describing important impressions from medical images through a new weight function.
- This paper introduces a Label Discriminator (LD), which helps classify medical images with more fine-grained labels.
- Extensive experiments over two known public datasets, the IU X-ray dataset and the PEIR Gross dataset, have demonstrated the effectiveness of the SVAML framework.

The remainder of this paper is organized as follows. The Related Work section rigorously compares this work in the context of the literature. The Methodology section will introduce in details the proposed SVAML framework. The Experiments section discusses the empirical studies over twp real-world datasets. The Conclusions section summarizes the paper.

Related Work

Medical image report generation usually consists of two stages. The one stage is to generate image subtitles, and the second one is to combine subtitles and generate reports in sentences (Yang et al. 2021). In the first stage, medical images are processed to generate feature words. Typically, CNN can be applied to detect target images and generate category words. However, CNN requires a huge amount of training data and its training efficiency is questionable. Meanwhile, because not many category words may be generated, it affects on subsequent sentence generation. In recent years, the Transformer architecture, originally from the NLP field, has been applied to image processing and obtained satisfactory results (Dosovitskiy et al. 2021), even surpassing convolution-based structures in some tasks. Based on the Transformer architecture, CNNs are not an indispensable component for image classification tasks any longer. For example, the Vision Transformer (ViT) (Dosovitskiy et al. 2021) proposes a pure Transformer architecture applied directly to sequences of embeddings of image patches split from the original images. When ViT combines with CNNs, a resulting convolutional-like ViT architecture called ConViT (d'Ascoli et al. 2021) exhibits sample-efficient learning and performance improvements. Since medical images in a particular domain, for example for glaucoma CT scans, may be characterized by a limited availability of samples. In our research, we decided to adopt the ConViT to learn image features.

Tang et al. (Tang et al. 2022) develops a H-Decoder to extract semantic features from medical images, centered by a pair of LSTM networks. The initial LSTM is dedicated to encoding tag features, while the subsequent LSTM is designed to facilitate sentence generation. The second LSTM network actually execute twice to encoding, to produce a pair of coherent paragraphs that collectively constitute a comprehensive medical report. In this work, we construct our decoder in the SVAML by adopting the principle of the H-Decoder.

In contrast to existing works, our proposed SVAML framework aims to automate medical image report generation. While we leveraged ConViT as encoder for image feature extraction and H-Decoder as decoder for sentence creation, SVAML introduces two new modules to enhance medical image report generation: a Double-Weighted Multi-Head Attention mechanism with a new weight function to learn patterns of how to focus on describing important impressions throuth the new weight function from medical images; and a 'Label Discriminator' to learn intricate features to support more sensitive multi-label classification.

SVAML Methodology

Fig. 1 illustrates the overview of our proposed framework, Semantic and Visual Attention-Driven Multi-LSTM Network (SVAML), tailored for automatic medical image report generation. As shown in Fig. 1, SVAML employs an encoder-decoder framework, comprising four modules: (1) Image Feature Extraction and Processing, (2) Double-Weighted Multi-Head Attention Encoder, (3) H-Decoder, and (4) Lable Discriminator. The following sections will explain the detailed designs of each module sequentially.

(1) Image Feature Extraction and Processing We utilize the ConViT model as part of the encoder for SVAML. Due

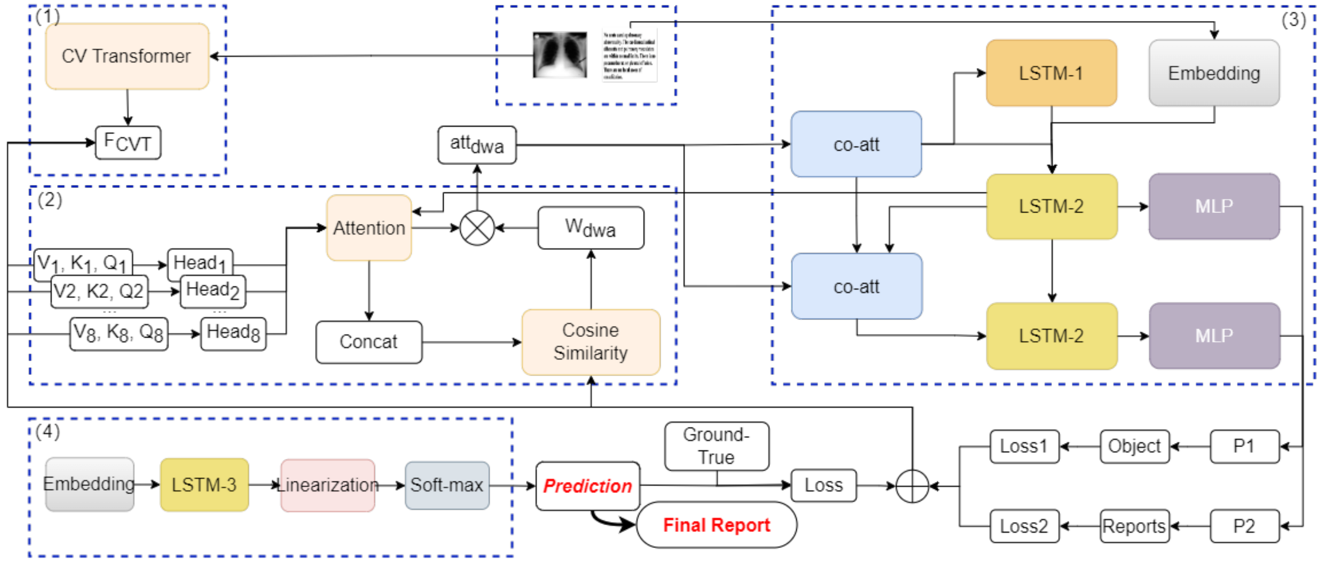

Figure 1: Overall architecture of SVAML. It consists of four modules: (1) Image feature extraction based on ConViT; (2) Double-Weighted Multi-Head Attention; (3) H-Deconder; (4) Label Discriminator. The image will first be sent to $CVTransformer$ for processing, generating F_{CVT}, and then input to the dual-weighted multi-head attention mechanism we designed. At the same time, $Label$ on the H-Deconder side will also be input to $LSTM$ at the same time. The two are processed in these two parts respectively, during which they will share a part of the weight. After that, $LabelDiscriminator$ will process both, calculate the loss function, and get the final report.

to the fact that medical images usually have limited samples, we pre-trained the module on the ImageNet dataset, omitting its final classification layer, for the extraction of 512-dimensional visual features (d'Ascoli et al. 2021). Based on the pre-trained model, we inherit from the SVEH architecture (Tang et al. 2022), incorporate an Image Feature Encoding (IFE) module to concurrently assimilate and encode both macroscopic and microscopic image attributes. As shown in Fig. 1, we made some modulation for the module, chaining the IFE behind ConViT to form a new CV Transformer. The image I is first processed by ConViT to obtain a feature map, which will serve as the input for IFE to obtain the final visual features F_{CVT}.

(2) Double-Weighted Multi-Head Attention We design a tailored attention module. The base attention model contains one head. It then splits one Q, K and V into N parts (N>1), each of which focusing on different aspects to enrich model features, and results in multi-head attention. N is the number of heads (Vaswani et al. 2017).

If applied conventional multi-attention mechanisms, their aggregation of N attention heads is achieved through straightforward concatenation or addition operations, thereby assigning equal significance to each head toward the generation of final reports. However, some areas that require "special attention," such as keywords, target objects, etc., require more attention, that is, they require greater weight.

Drawing upon this perceptual insight, our research introduces the Multi-Head Attention that is designed to more accurately emulate the selective attentional processes observed in human cognition by differentially weighting the significance of various attention heads, thereby aligning the mechanism more closely with the nuanced manner in which humans process visual stimuli. $head_a^i$ is the output, being processed by different heads i, and it will be mapped linearly by a group of weights w_a. At the beginning of each training iteration, $w_{a(i)}$ (i means the iteration of training) is calculated by $w_{a(i-1)}$, multiplying softmax function. At the first epoch of model training, each head share equal importance. Due to it, $w_{a(0)}$ is a straight vector (all values are 1.).And we use BN, a normalization factor, to multipy $w_{a(i)}$, which is used to keep its consistency throughout the whole training process, shown as $w_{a(i)} = softmax(w_{a(i-1)})$. Based on it, the output of attention module with single weight can be shown as $head_{wa(i)} = w_{a(i)} * head_{a(i)}$.

For multi-head, it becomes apparent that certain heads, which are more pertinent to report generation, acquire significantly higher weights following the training phase. This results in a pronounced emphasis on these heads in the overall process. Conversely, other heads assume more nuanced roles, contributing in a less pronounced, yet still integral, manner to the comprehensive functionality of the model. To enhance this advantage, we introduce multi-headed attention. For each step, the head whose W_a is the highest will be chosen as the base for the next step. We use Cosine Weight $cos(i)^j$ to represent the cosine similarity between head i and base at the jth iteration. For each batch, the computation of the second weight $w^j_{cos(i)}$ involves aggregating the cosine similarities corresponding to the same attention head. This method ensures a balanced and reasonable approach to de-

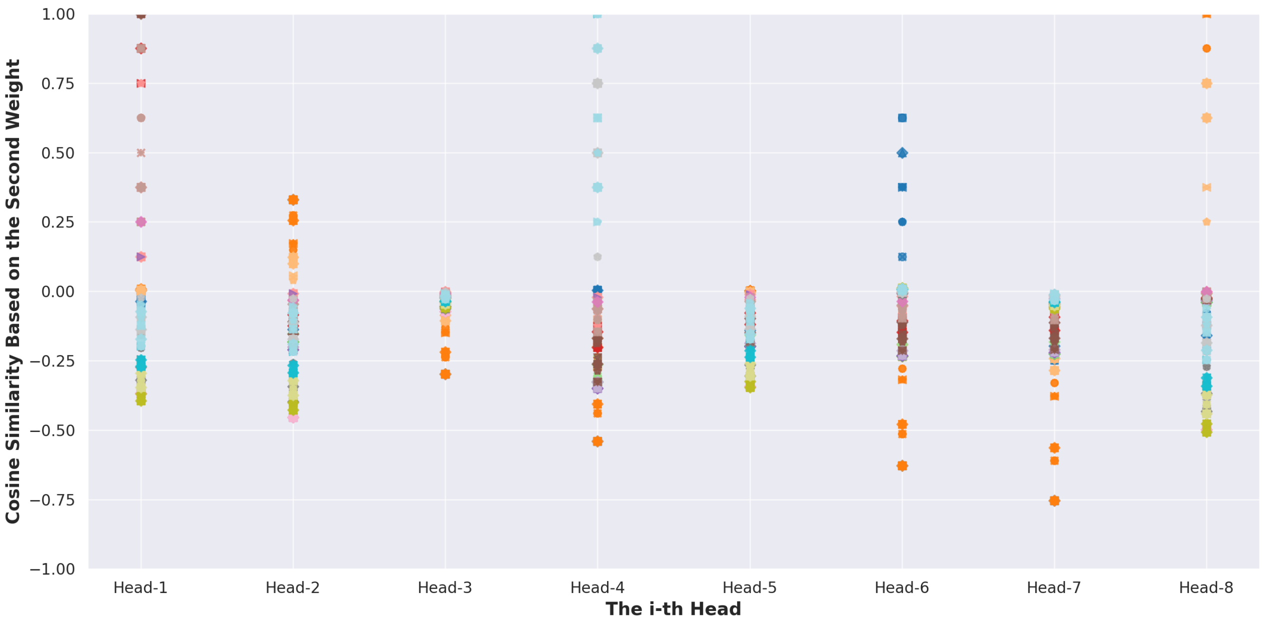

Figure 2: The value of cosine similarity is related to the distribution of the values of different heads. Due to the large amount of data, we only used 5000 Cosine Similarities of all for each head to describe its distribution characteristics. The distribution pattern of all data itself is almost the same as this.

termining the secondary weighting factor within the multi-head attention framework.

$$cos(i)^j = cos(head^j_{a(i-1)}, head^{base}_{a(i-1)}) \tag{1a}$$

$$w^j_{cos(i)} = \frac{\sum_{k \in i} cos(i)^j_k}{N} \tag{1b}$$

$$w^j_{dwa(i)} = w^j_{a(i)} * (1 - w^j_{cos(i)}) \tag{1c}$$

$$head^j_{dwa(i)} = w^j_{dwa(i)} * head^j_{a(i)} \tag{1d}$$

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

As shown in Fig. 2, we calculate the average cosine similarity during multiple training phases. It can be seen that the cosine similarity between the same head is 1, which is the diagonal value of the matrix. However, due to the similarity of most of the labels in the dataset itself and the image itself, the similarities calculated by different heads are very close to 1, with only a few negative values. This is very close to the label sentence (Ground-Truth) itself, that is, the sentence structure and wording, and is very consistent with the different characteristics of only the keywords description of disease. Based on the value distribution of cosine similarity, we design a weight function of $1 - w^j_{cos(i)}$ to balance the weight features. Be specific, we hope our model to be able to remember negative samples, i.e., diseased samples, more accurately. The more similar the normal sample specimens are, the cosine similarity will be close to 1, $1 - w^j_{cos(i)}$ will be close to 0, and the weight itself will not be too much amplified. The closer the negative sample is to -1, the value is amplified, that is, the closer it is to 2. In this way, the weight of the negative sample is strengthened. In other words, the designed model will focus more on learning from those negative samples.

(3) H-Decoder As shown in Fig. 1, H-Decoder consists of two LSTMs and the second LSTM (LSTM-2) decodes twice (Tang et al. 2022). Unlike the Look Back (LB) method (Qin et al. 2019), in this architecture, shown in Figure 2, LSTM-1 is exclusively employed for the encoding of tag features, serving a pivotal role in the initial processing stage. Concurrently, LSTM-2 is utilized for the generation of sentences, indicating a functional distinction between two-LSTM networks. This bifurcation in roles allows for a more specialized and efficient handling of distinct tasks within the overall computational process. The dimensions for both word embedding and the hidden states of all LSTM networks are uniformly set at 512.

(4) Label Discriminator Medical imaging datasets frequently incorporate explicit tags that denote the type of disease and other critical information. Contemporary computational models utilize such tags within a multi-label classification framework to facilitate tag prediction, subsequently leveraging these predictions in the generation of diagnostic reports. This approach hinges on the accuracy of tag prediction. However, the efficacy of image classification models is often compromised by the limited size of most medical datasets, which simultaneously contain a vast array of tags.

Consequently, the reliability of report generation, being intrinsically linked to the precision of tag prediction, is undermined. Inaccurate tag predictions not only render the reports ineffective but can also detrimentally impact the overall performance of the model.

Tags can be seen as key words of reports. They are just like symbols that can be used for identification. Inspired by this relationship, we propose a novel method to utilize symbol effectively. Because of it, training process will be more efficientive. The architecture of this model is distinguished by its inclusion of a multi-label classification module. Uniquely, in a deviation from established methodologies, this module processes generated textual reports as its primary input, rather than direct image data. At its core, the module integrates a LSTM network, explicitly designed for the prediction of tags. Complementing this is a dedicated classification layer, which together with the LSTM, whose dimensions for both word embedding and the hidden states are 512, synergizes to effectively categorize and analyze the content of the reports. This innovative approach allows for a more nuanced interpretation of the textual data, thereby augmenting the precision and reliability of the classification outcomes in the context of the model's broader application. Embedded reports R_e are sent into the prediction LSTM, and it outputs predict symbol. Then, we flatten s with a linear mapping layer and gain probability of each symbol with softmax. This symbol participates in the training process by adding its loss function, $loss_T$, in backward propagation. All relationships can be shown as below:

$$s^{'} = LSTM(R_e) \tag{2a}$$

$$s = softmax(W_t * s^{'} + b_t) \tag{2b}$$

$$loss_{t1} = \sum\nolimits_i tag_i * log(\frac{e^{s_i}}{1 + e^{s_i}}) \tag{2c}$$

$$loss_{t2} = (1 - tag_i) * log(\frac{1}{1 + e^{s_i}}) \tag{2d}$$

$$loss_T = -\frac{1}{C} loss_{t1} + loss_{t2} \tag{2e}$$

$$loss = loss_1 + \eta * loss_2 + \lambda * loss_T \tag{2f}$$

where W_t and b_t are trainable parameters used to flatten $s^{'}$; tag represents the true 'symbol (tags)'; $i \in \{0, ..., n-1\}$, $tag_i \in \{0, 1\}$; 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) LSTM-2 in H-Decoder and true captions. The second LSTM-2 generation is confined by an adjustable parameter $\eta \in (0, 1]$. Due to the reason that $loss_T$ is merely one-tenth of $loss_1$ and $loss_2$, $loss_T$ needs to be amplified to balance. Given that $loss_t$ does not reflect to reports quality like $loss_1$ and $loss_2$, so we set its coefficient λ at 5. And $loss_T$ is the MultiLabel-SoftMarginLoss[1].

Experiments

Datasets and Evaluation Metrics

Datasets Two widely-used datasets, **IU X-ray** and **PEIR Gross** are used in this paper. **IU X-ray** is a chest X-ray collection selected from the Indiana Network for Patient Care by researchers from Indiana University[2]. **PEIR Gross** is the Gross sub-collextion of the Pathology Education Informational Resource (PEIR) digital library[3]. For **IU X-ray**, we adopt 6730 image-caption pairs, which has been processed. For **PEIR Gross**, We have used 7442 image-caption pairs. For both two datasets preparation, please refer to the survey (V. Kougia 2019) and the journal (Demner-Fushman et al. 2016).

Evaluation Metrics The optimization of our model is carried out using the ADAM algorithm (Kingma and Ba 2015), with a learning rate set at 0.0004. Following the insights (Tang et al. 2022), we adjust the value of η to 0.5 to optimize performance. For the evaluation process, we implement a 10-fold cross-validation technique on both datasets. In this schema, each fold comprises a set of 500 distinct, non-overlapping images, randomly selected to ensure a representative sample. The remaining images in the dataset are consistently utilized as the training set. This methodology ensures a robust validation of the model's performance across a diverse range of data samples. During the evaluation phase, a beam search strategy is employed. And three widely-used metrics are used to evaluate our work : BLEU (Papineni et al. 2002), ROUGEL (Lin and Och 2004), and CIDEr (Vedantam, Zitnick, and Parikh 2015). For the computation of metrics, we utilize a widely-adopted image captioning tool[4].

Experimental Results

Comparison to Advanced Models As shown in Table 1, we compare many well-performed previous works and our model obtains state-of-the-art results on both validation datasets. Regardless of whether the sentences are long or short, our model achieves the highest accuracy on IU X-ray. For PEIR Gross, we got the highest ones on BLEU-1, BLEU-2, BLEU-3 and ROUGEL, and the sencond highest ones on BLEU-5 and CIDEr.

We also show some comparison reports of different models generated corresponding to images, as shown in Figure 3. By comparison with other models, our model generates more concise reports. Compared with the positive samples, the difference between the reports and Ground-Truth is not very big. Because of $1 - w^{j}_{cos(i)}$, the model itself will not have a high weight when learning normal samples. Compared with pathological samples, because their weights are amplified, the model "pays more attention" to these keywords and differences. Therefore, the final generated report is more accurate in capturing pathological nouns and will not be much different from the original Ground-Truth. The

[1] MultiLabel- SoftMarginLoss: https://pytorch.org/docs

[2] https://openi.nlm.nih.gov

[3] https://peir.path.uab.edu/library/

[4] https://github.com/Maluuba/nlg-eval

Model	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGEL	CIDEr
cnn-rnn (Vinyals et al. 2015)	IU X-ray	38.1	29.0	22.7	16.0	39.6	31.2
co-att (Jing, Xie, and Xing 2018)		46.2	33.1	24.2	17.8	40.5	40.8
nearest-neighbor (V. Kougia 2019)		28.1	15.2	9.1	5.7	20.9	-
KERP (Li et al. 2019)		48.2	32.5	22.6	16.2	33.9	28.0
MvH (Yuan et al. 2019)		43.6	31.2	22.9	17.0	37.2	32.8
A3FH (Xie et al. 2019)		44.3	33.7	23.6	18.1	34.7	-
JE-TriNet (Yang et al. 2021)		47.8	34.4	24.8	18.0	39.8	43.9
TransGen (Jia et al. 2021)		46.1	28.5	19.6	14.5	36.7	-
PPKED (Liu et al. 2021)		48.3	31.5	22.4	16.8	37.6	35.1
SVEH (Tang et al. 2022)		50.8	35.6	25.9	19.1	40.8	41.5
AENSI (Huang et al. 2023)		54.2	36.4	26.7	19.8	43.3	46.4
SVAML (Ours)		**55.7**	**37.7**	**28.1**	**21.3**	**43.5**	**47.5**
co-att (Jing, Xie, and Xing 2018)	PEIR Gross	30.0	21.8	16.5	11.3	27.9	32.9
nearest-neighbor (V. Kougia 2019)		34.6	26.2	20.6	15.6	34.7	-
SVEH (Tang et al. 2022)		46.6	32.3	23.3	16.9	37.4	26.9
AENSI (Huang et al. 2023)		44.2	31.5	22.6	17.4	43.5	28.2
SVAML (Ours)		**46.7**	**33.3**	**24.1**	**17.2**	**44.6**	**31.2**

Table 1: Comparison of Proposed Methods with state-of-the-art Methods on the IU X-ray and PEIR Gross. Red means that except for our method, the result is the highest. **-** means that we did not get test results through experiments, and consulting the original article did not provide results. (×100%)

Model	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGEL	CIDEr
Baseline	IU X-ray	51.0±1.4	31.8±1.7	24.9±2.3	16.8±2.6	38.9±2.1	42.4±3.4
+Single-Weight		51.1±1.7	31.9±1.3	25.3±2.0	16.9±3.4	39.1±2.5	43.3±2.6
+Quadruple-Weight		52.4±2.1	32.6±1.9	25.9±1.1	18.7±2.1	39.5±1.7	43.2±3.1
+Lable Discriminator		52.5±2.4	33.4±2.3	26.1±1.4	19.4±1.1	40.0±2.9	44.1±2.4
+SVAML (All)		53.2±2.5	34.0±3.7	26.5±1.6	19.8±1.5	40.1±3.4	44.8±2.7
Baseline	PEIR Gross	42.2±1.7	28.1±1.7	19.4±2.2	14.2±1.2	41.6±2.5	24.2±1.7
+Single-Weight		42.0±2.9	28.5±2.1	19.2±2.7	14.2±1.3	41.1±2.5	23.3±4.7
+Quadruple-Weight		42.7±2.1	29.1±2.7	20.1±2.4	13.8±2.7	41.8±3.1	24.9±3.1
+Lable Discriminator		42.8±2.7	29.2±3.1	20.7±3.0	14.9±1.9	41.9±3.5	21.9±7.2
+SVAML (All)		43.2±3.5	29.9±3.4	20.9±3.2	15.1±2.1	42.4±4.2	25.3±5.9

Table 2: Ablation Study of Key Structures on IU X-ray Dataset and PEIR Gross: The baseline model comprises a Vision Transformer (ViT) encoder, a Hierarchical Decoder (H-Decoder), and a multi-head attention module. (×100%)

reports generated by JE-TriNet ((Yang et al. 2021)), whether for normal conditions or pathological conditions, have relatively long and complicated sentences. This will have some impact on the efficiency of people reading reports and extracting useful pathological information.

Meanwhile, as shown in Fig. 4, $Loss$, $Loss_1$, $Loss_2$ and $Loss_T$ all tend to stabilize after the fourth epoch. During the training process, the model itself focuses on pathological samples and does not pay much ”attention” to normal samples. Because of this, training is faster and declines faster.

Ablation Experiment Also, we have tested baselines of our model. They are a ViT encoder, H-Decoder and traditional multi-head attention module, based on the IU X-ray dataset and PEIR Gross, as shown in Table 2. The comparative analysis reveals a consistent trend of enhancement with the incorporation of our proposed modules. Most notably, our fully developed model, SVAML, achieves the most superior results. Additionally, the comparison between DW-MHA and Lable Discriminator indicates a close alignment in their effectiveness. Furthermore, the incremental improvements observed in the top three rows substantiate the impact of the dual heads in DW-MHA. This is particularly evident as the addition of three weights leads to a more pronounced improvement than the inclusion of a single weight alone, highlighting the superiority of dual-head performance over a single-head approach.

Conclusion

In this paper, we present two innovative solutions to address the predominant challenges in current medical report generation models. Firstly, we introduce a double-weighted multi-head attention mechanism with a new weight function, which enhances the model's ability to concentrate on the most significant segments of images during the generation process. Secondly, we propose the Lable Discriminator (LD), which optimally utilizes tag information within the constraints of limited training samples. Our extensive and comprehensive experiments across both radiology and pathology datasets validate the effectiveness of our meth-

Image	Label	Ground-True	JE-TriNet	AENSI	SVAML
	normal	No acute cardiopulmonary abnormality. The cardiomediastinal 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 process. PA and lateral views of the chest provided. The lungs are clear without focal consolidation. The cardiomediastinal silhouette is normal. No pleural effusion or pneumothorax is seen. No free air below the right hemidiaphragm is seen.	No acute cardiopulmonary abnormality.The lungs are clear. There is no focal air space opacity. No pleural effusion orpneumothorax. The cardiomediastinal silhouette is normal.	No acute cardiopulmonary abnormality. The cardiomediastinal silhouette and pulmonary vasculature are normal. There is no pneumothorax or pleural effusion. No focal airspace.
	spine degenerative	No acute cardiopulmonary process. Normal heart size and mediastina contours. Lungs are clear. There is no pneumothorax or pleural effusion. Degenerative changes are seen in the spine.	Increased interstitial markings bilaterally may be due to chronic interstitial lung disease. The heartis mildly enlarged. The heart is mildly enlarged. There is no pleural effusion or pneumothorax. There is nopleural effusion or pneumothorax.	No active disease. Both lungs are clear and expanded. Heart and mediastinum within normal limits. Degenerative changes in the thoracic spine.	No acute cardiopulmonary process. Heart size and mediastina contours are normal. Lungs are clear. There is no pneumothorax or pleural effusion. Degenerative changes in the spine.

Figure 3: Examples of medical reports generated by SVAML and other models on IU X-ray

ods. Furthermore, when benchmarked against current state-of-the-art models, our approach demonstrates equivalent or superior performance.

Future Work

We also found some problems and interesting places when we were doing experiments.

- We found that when our model was tested on the PEIR Gross data set, the evaluation matrix CIDEr fluctuated particularly large. Its experimental results are not as good as the previous basic model. Because of this, in future research, we will analyze the applicability of our model itself and the characteristics of the data set, and improve our algorithm based on the paper (Tao Tu 2023).
- The ConViT model (d'Ascoli et al. 2021) still has some shortcomings when extracting local features: it cannot capture small edges, requires a high-precision data set, and requires powerful GPU resources during training. Based on the above defects, we will optimize this part, or redesign a Transformer algorithm, or use the latest EfficientViT for experiments (Liu et al. 2023; Cai, Gan, and Han 2022).
- We will redesign a weight function in future research. Because a large number of images and their label descriptions are very similar, the pathological causes are also very similar. Because of this, we will focus on strengthening our focus on 'differences'. Remembering only some particularly distinctive characteristics (pathological), not necessarily some general (normal) characteristics will make the model more efficient. We will further improve this point and make it adaptive to changes in weights, so that sample features can be learned more accurately.
- We will combine some emerging fields, such as AIoT (Baker and Xiang 2023; Chen et al. 2022), edge computing (Lin et al. 2023; Liu et al. 2019), etc. Finally, we will develop the model into an industrial-grade AI that can be applied on the ground.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. 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 (CVPR)*, 6077–6086. Salt Lake City, UT, USA: IEEE.

Baker, S.; and Xiang, W. 2023. Artificial Intelligence of Things for Smarter Healthcare: A Survey of Advancements, Challenges, and Opportunities. *IEEE Communications Surveys & Tutorials*, 25(2): 1261–1293.

Cai, H.; Gan, C.; and Han, S. 2022. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv:2205.14756.

Chen, J.; Zheng, Y.; Liang, Y.; Zhan, Z.; Jiang, M.; Zhang, X.; da Silva, D. S.; Wu, W.; and Albuquerque, V. H. C. d. 2022. Edge2Analysis: A Novel AIoT Platform for Atrial Fibrillation Recognition and Detection. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 5772–5782.

Chen, Z.; Wu, M.; Cui, W.; Liu, C.; and Li, X. 2020. An Attention Based CNN-LSTM Approach for Sleep-Wake Detection With Heterogeneous Sensors. *IEEE Journal of Biomedical and Health Informatics*, 25: 3270–3277.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23: 304–310.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021.

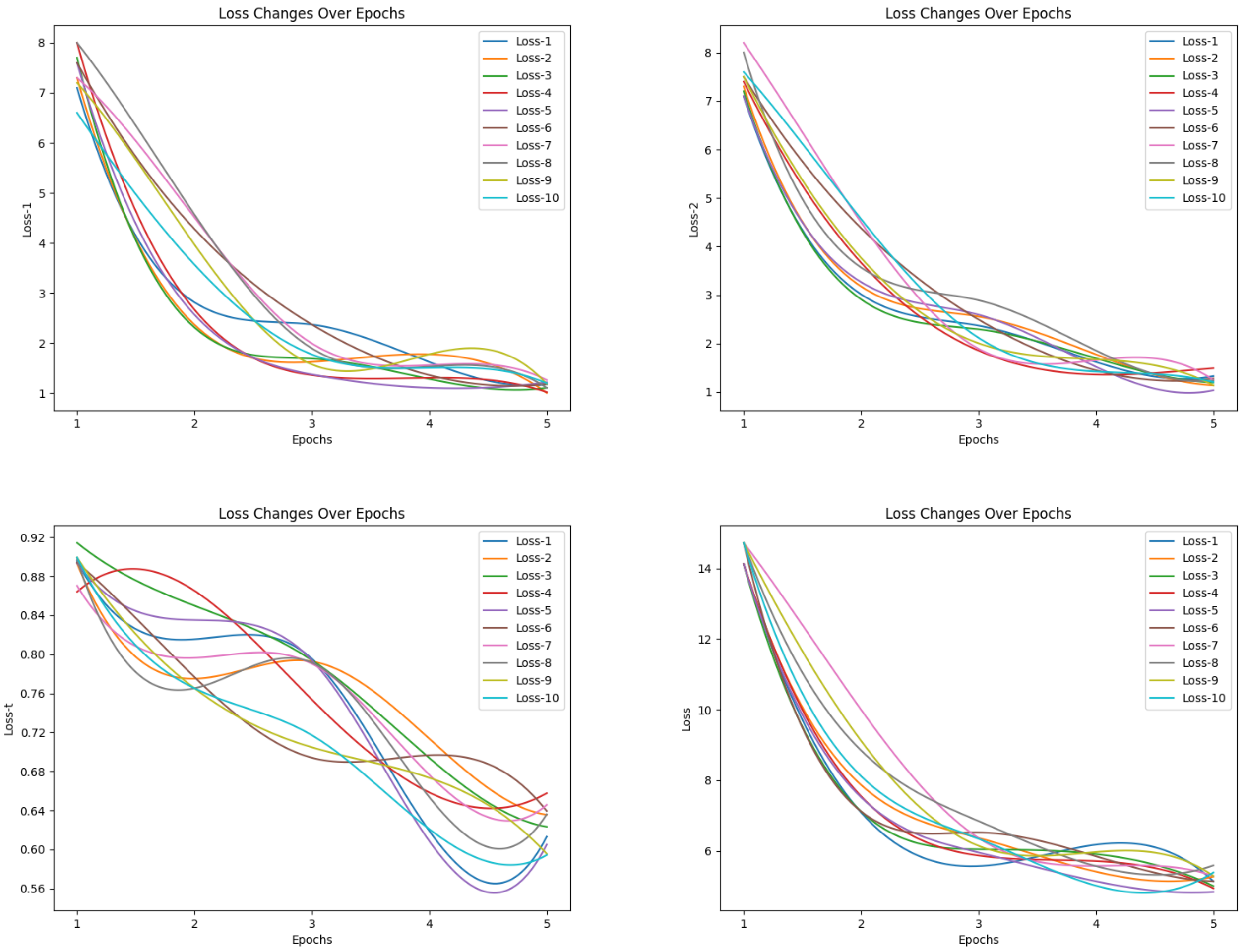

Figure 4: Their values of $Loss$, $Loss_1$, $Loss_2$ and $Loss_T$ change as epoch increases. $Loss\text{-}i$ on the way represents the i-th training. We use 10-fold cross-validation.

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of The 9th International Conference on Learning Representations (ICLR)*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2286–2296. Vienna, Austria.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, 15–29. Berlin, Heidelberg: Springer-Verlag.

Gavrilyuk, K.; Sanford, R.; Javan, M.; and Snoek, C. G. M. 2020. Actor-Transformers for Group Activity Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 836–845. Seattle, WA, USA: IEEE.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9: 1735–1780.

Huang, C.; Lin, Y.; Tang, Q.; Wang, H.; and Yu, Y. 2023. Attention Enhanced Network with Semantic Inspector for Medical Report Generation. In *The 35th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. Atlanta, GA, USA: IEEE.

Jia, X.; Xiong, Y.; Zhang, J.; Zhang, Y.; Suzanne, B.; Zhu, Y.; and Tang, C. 2021. Radiology Report Generation for Rare Diseases via Few-shot Transformer. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1347–1352. Houston, TX, USA: IEEE.

Jing, B.; Xie, P.; and Xing, E. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2577–2586. Melbourne, Australia: ACL.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2019.

Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. arXiv:1903.10122.

Liang, J.; Homayounfar, N.; Ma, W.-C.; Xiong, Y.; Hu, R.; and Urtasun, R. 2020. PolyTransform: Deep Polygon Transformer for Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9128–9137. Seattle, WA, USA: IEEE.

Lin, B.-S.; Peng, C.-W.; Lee, I.-J.; Hsu, H.-K.; and Lin, B.-S. 2023. System Based on Artificial Intelligence Edge Computing for Detecting Bedside Falls and Sleep Posture. *IEEE Journal of Biomedical and Health Informatics*, 27(7): 3549–3558.

Lin, C.-Y.; and Och, F. J. 2004. 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)*, 605–612. Barcelona, Spain: ACL.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13748–13757. Nashville, TN, USA: IEEE.

Liu, H.; Yao, X.; Yang, T.; and Ning, H. 2019. Cooperative Privacy Preservation for Wearable Devices in Hybrid Computing-Based Smart Health. *IEEE Internet of Things Journal*, 6(2): 1352–1362.

Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; and Yuan, Y. 2023. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Maitre, J.; Bouchard, K.; and Gaboury, S. 2020. Fall Detection With UWB Radars and CNN-LSTM Architecture. *IEEE Journal of Biomedical and Health Informatics*, 25: 1273–1283.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. Philadelphia, PA, USA: ACL.

Qin, Y.; Du, J.; Zhang, Y.; and Lu, H. 2019. Look Back and Predict Forward in Image Captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8359–8367. Long Beach, CA, USA: IEEE.

Tang, Q.; Yu, Y.; Feng, X.; and Peng, C. 2022. Semantic and Visual Enrichment Hierarchical Network for Medical Image Report Generation. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, 738–743. Hangzhou, China: IEEE.

Tao Tu, e., Shekoofeh Azizi. 2023. Towards Generalist Biomedical AI. arXiv:2307.14334.

V. Kougia, e. 2019. A Survey on Biomedical Image Captioning. arXiv:1905.13302.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Łukasz Kaiser; and Polosukhin, I. 2017. Attention Is All You Need. In *Proceedings of the Thirty-first Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA: Curran Associates Inc.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575. Boston, MA, USA: IEEE.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164. Boston, MA, USA: IEEE.

Wang, Z.; Han, H.; Wang, L.; Li, X.; and Zhou, L. 2022. Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach. *IEEE Transactions on Medical Imaging*, 41: 2803–2813.

Xie, X.; Xiong, Y.; Yu, P. S.; Li, K.; Zhang, S.; and Zhu, Y. 2019. Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation. In *International Conference on Database Systems for Advanced Applications (DASFAA)*, 448–452. Springer.

Yang, Y.; Yu, J.; Zhang, J.; Han, W.; Jiang, H.; and Huang, Q. 2021. Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation. *IEEE Transactions on Multimedia*, 25: 167–178.

Yuan, J.; Liao, H.; Luo, R.; and Luo, J. 2019. Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment. arXiv:1907.09085.

Zheng, R.; Wang, Q.; Lv, S.; Li, C.; Wang, C.; Chen, W.; and Wang, H. 2022a. Automatic Liver Tumor Segmentation on Dynamic Contrast Enhanced MRI Using 4D Information: Deep Learning Model Based on 3D Convolution and Convolutional LSTM. *IEEE Transactions on Medical Imaging*, 41: 2965–2976.

Zheng, Y.; Gindra, R. H.; Green, E. J.; Burks, E. J.; Betke, M.; Beane, J. E.; and Kolachalama, V. B. 2022b. A Graph-Transformer for Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 41: 3003–3015.

Zhou, X.; Li, Y.; and Liang, W. 2020. CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18: 912–921.