

Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach

Zhanyu Wang^{ID}, Hongwei Han^{ID}, Lei Wang^{ID}, *Senior Member, IEEE*, Xiu Li^{ID}, *Senior Member, IEEE*, and Luping Zhou^{ID}, *Senior Member, IEEE*

Abstract—Automated radiographic report generation is challenging in at least two aspects. First, medical images are very similar to each other and the visual differences of clinic importance are often fine-grained. Second, the disease-related words may be submerged by many similar sentences describing the common content of the images, causing the abnormal to be misinterpreted as the normal in the worst case. To tackle these challenges, this paper proposes a pure transformer-based framework to jointly enforce better visual-textual alignment, multi-label diagnostic classification, and word importance weighting, to facilitate report generation. To the best of our knowledge, this is the first pure transformer-based framework for medical report generation, which enjoys the capacity of transformer in learning long range dependencies for both image regions and sentence words. Specifically, for the first challenge, we design a novel mechanism to embed an auxiliary image-text matching objective into the transformer's encoder-decoder structure, so that better correlated image and text features could be learned to help a report to discriminate similar images. For the second challenge, we integrate an additional multi-label classification task into our framework to guide the model in making correct diagnostic predictions. Also, a term-weighting scheme is proposed to reflect the importance of words for training so that our model would not miss key discriminative information. Our work achieves promising performance over the state-of-the-arts on two benchmark datasets, including the largest dataset MIMIC-CXR.

Index Terms—Medical report generation, image caption, transformer, image-text matching.

I. INTRODUCTION

AUTOMATED generation of diagnostic reports upon medical images is a very challenging task, which bridges

Manuscript received 13 February 2022; revised 28 March 2022; accepted 23 April 2022. Date of publication 4 May 2022; date of current version 30 September 2022. The work of Xiu Li was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108303 and in part by NSFC under Grant 41876098. (Corresponding author: Luping Zhou.)

Zhanyu Wang and Luping Zhou are with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: zwan0839@uni.sydney.edu.au; luping.zhou@sydney.edu.au).

Hongwei Han and Xiu Li are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Beijing 100084, China (e-mail: hhw20@mails.tsinghua.edu.cn; li.xiu@sz.tsinghua.edu.cn).

Lei Wang is with the School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: leiw@uow.edu.au).

Digital Object Identifier 10.1109/TMI.2022.3171661

visual and linguistic information to generate free-form text describing the findings on a medical image. The disproportional growth of radiographic images against trained readers has exerted massive pressure on radiologists to examine medical images and write reports. This has resulted in a large number of backlog of unreported radiological examinations, particularly in public hospitals, and increased the likelihood of errors in medical reports. Also, the quality of the reports depends on the expertise and experience of individual radiologists. Therefore, the automation of such process is in high demand and becoming a prominent and attractive research topic.

Medical report generation roots in a more general research topic of image captioning. A comprehensive review is given in Section II. The state-of-the-art image captioning methods are dominated by deep learning models. They mostly follow a basic encoder-decoder structure proposed in [1], [2], where Convolution Neural Network (CNN) serves as a visual encoder to extract the visual features from images and the Recurrent Neural Network (RNN, normally LSTM [3]) as a text decoder to convert the visual features to text output. Various attention mechanisms have also been developed to guide the gaze of objects when generating captions [2], [4]–[7]. Recently the choice of text decoder is shifting from RNNs to transformers [8], due to their strong capacity of learning long range dependencies. Although transforms are increasingly used in vision tasks such as image classification [9] and object detection [10], pure transformer-based visual encoder (without CNN) has not been well studied for image captioning.

Despite their achievements, it is known that current image captioning methods usually generate overly rigid reports which do not work well on discriminating similar images and tend to repeat the phrases that frequently appear in the training set [11]. This makes them difficult to be effectively applied to radiographic report generation due to two challenges. First (Challenge I), medical images are very similar to each other and clinic importance is often reflected via fine-grained visual differences. Second (Challenge II), the disease-related words may be submerged by many similar sentences describing the common content of the images. This could cause the abnormal to be misinterpreted as the normal in the worst case. Existing medical report generation methods have made extra efforts,

such as training additional classifiers to predict disease labels or medical tags [12], [13] or utilising knowledge graph as prior [14], to improve the report quality. In addition, since medical reports are long narratives of multiple sentences, efforts have also been spent in developing models, such as hierarchical LSTM [12], [15], for long paragraph generation.

To advance radiographic report generation, we propose a multi-criteria supervised transformer in this paper. The whole model is built purely upon transformer without the need of CNN. Different from common image captioners using CNN to extract visual features, our transformer based visual encoder circumvents the limited receptive field of CNNs, so that long-range visual dependency could be effectively learned. As for our text decoder, a memory-augmented transformer replaces the commonly used hierarchical LSTM for long paragraph generation, which could fully explore word relationship in long range. Moreover, to respond to the two challenges mentioned above, we bring forward a multi-criteria supervision framework to better train our transformer for report generation. Specifically, for Challenge I, an objective of image-text matching is tactically proposed to force strongly correlated visual and text features to be learned, so that the generated reports could discriminate similar images. For Challenge II, a multi-label classification objective for medical tag prediction is employed to regularise report generation, as well as a proposed term weighting scheme that adjusts the importance of words for training. Our contributions are summarized as follows.

First, to the best of our knowledge, this is the first work that generates medical reports purely based on a transformer model. Transformer, as arguably the next generation scalable vision model, has made its first step in completely replacing CNN for image classification [9]. Our work demonstrates the possibility for a pure transformer model (under the proposed multi-criteria supervision) to achieve the state-of-the-art performance in medical report generation without using CNN, which has considerable referential importance to successive works. Compared with almost all existing medical report generation methods [12], [13], [16]–[18] that rely on a pretrained ResNet [19] or DenseNet [20] to extract image-level features for visual representation, our pure-transformer based approach extracts regional features without the need of generating region proposals, and explicitly considers the relationship among image regions for better visual representation. Please note that, although some CNN-based image captioning methods [21]–[24] use pretrained Fast-RCNN model [25] to detect object regions for regional features, they have to step back to image-level features for medical report generation since such datasets rarely have ground-truth object bounding-boxes to allow fine-tuning Faster-RCNN, while the latter is pretrained on natural images that do not share common object classes with medical images. Building upon the pure transformer architecture, our method could avoid this problem by circumventing the generation of region proposals when exploring regional features and relationships.

Second, we propose to utilise the image-text matching loss to enforce our model to learn strongly correlated image and text features for report generation. The image-text matching

loss explicitly measures the semantic similarity between a given image-text pair by classifying whether they are a true pair or not. In our approach, an image and its corresponding ground-truth report form a true (positive) pair, while an image and the report of any other images form a false (negative) pair. Differentiating whether a report matches the given image helps correlate the learned visual and textual features. Moreover, we design a novel mechanism that enables us to effectively embed the image-text matching task into the transformer framework with end-to-end training. Please note that this is not trivial to be achieved in a unified network since report generation and image-text matching have different ways of utilizing image and text features. With our design, we could effectively achieve this without the need to alter the transformer structure or introduce a separate branch for image-text matching, avoiding increasing the complexity of the network model while improving the quality of the generated reports.

Third, we bring in a term-weighting scheme and a multi-label classification objective to further supervise our model to accurately capture critical diagnostic information and generate reports. Specifically, the term-weighting scheme re-weights the importance of words in the reports, increasing the weight of discriminative terms that occur relatively rarely, while the multi-label classification is performed to predict the medical tags of the given image so that such information could be catered for in the feature learning.

Last, we extensively validate our model on two benchmarks, including the largest dataset MIMIC-CXR released recently. The results indicate that our framework outperforms multiple state-of-the-art methods in image captioning and medical report generation, exhibiting the great potential of transformer-based models for this challenging task.

II. RELATED WORK

A. Image Captioning

The state-of-the-art image captioning methods are deep learning based [1], [2], [5]–[7], [21]–[23], [26]–[29]. Most of them follow *Show-Tell* [1] to build upon the encoder-decoder structure, with a CNN based visual encoder to extract image features and a followed RNN based text decoder to convert visual features to text output. Object detection is usually employed to produce region level image features [25] and attention mechanisms have been developed to allow the text decoder to learn where and what it should attend in the image for caption generation [2], [5]–[7]. Recent years have witnessed the advance of both visual encoders and text decoders in image captioning. For visual encoders, most efforts are spent on digging visual relationships of the detected image regions [24], [30] via techniques such as graphical convolution networks or scene graph. For text decoders, the transformer model [8], which has achieved big success in natural language processing, starts to replace RNN as the text decoder [22] in image captioning tasks. Compared with RNN, transformer processes sentences as a whole without recursion, so that full word relationship beyond sequence could be better explored.

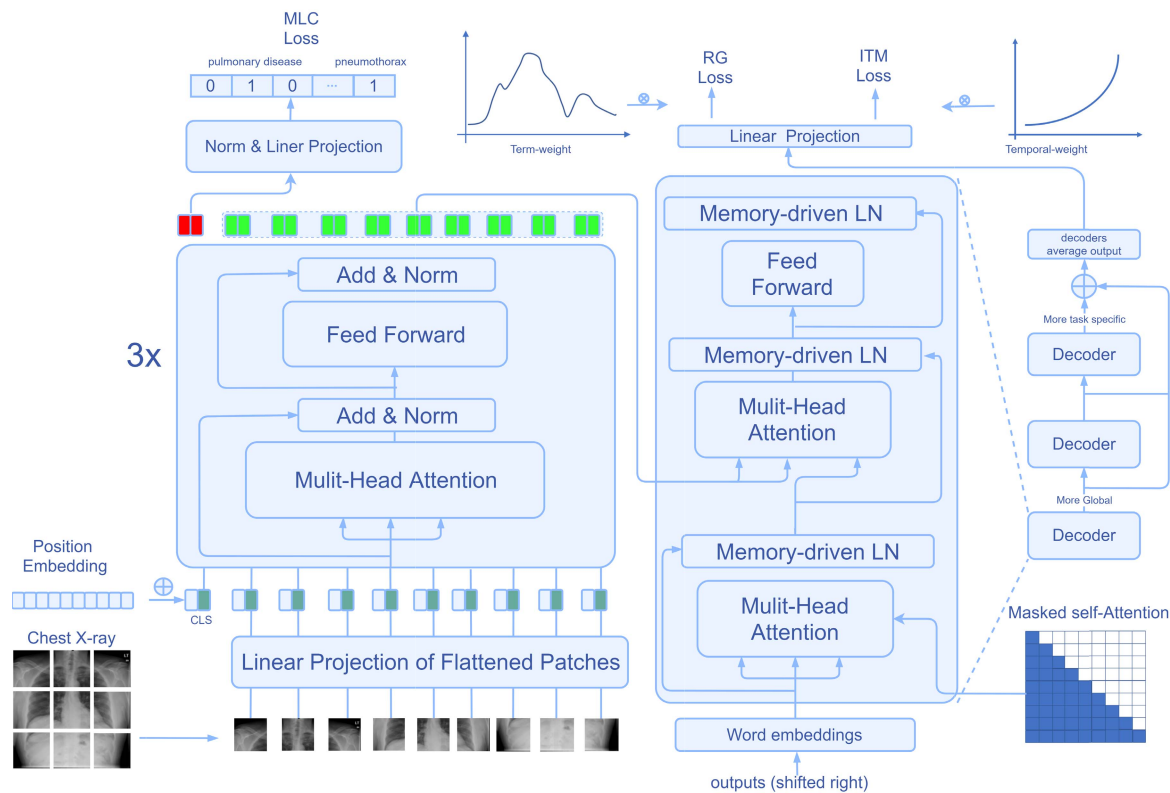


Fig. 1. An overview of the proposed framework which comprises of a vision transformer encoder and a memory augmented text transformer decoder. The whole model is supervised by three criteria: 1) a term weighting loss for the main task of report generation (RG loss), 2) an image-text matching loss with temporal-weighting (ITM loss), which facilitates report generation by enforcing the model to learn strongly correlated visual and textual features, and 3) a multi-label classification loss (MLC loss) that regularises the report generation by ensuring the model to make correct diagnostic prediction.

B. Medical Report Generation

Most research efforts in medical report generation are along two main directions. The first direction lies in developing deep learning models to better handle the long narrative nature of medical reports. Along this line, many existing methods employed the hierarchically structured LSTM network [12], [13], [16] to produce detailed text description of the findings from the radiographs. They comprise of a sentence LSTM to generate topics and a word LSTM to produce text output, which could be further trained with additional attention mechanisms [12]. In addition, a different network structure was proposed in [17], [18], which introduced a recurrent paragraph generative model, utilising the generated sentence to produce the next sentence. The second research direction studies how to leverage medical domain knowledge to guide report generation. This includes regularising report generation with additional disease or medical tag predictions [13] or injecting knowledge graph to incorporate prior [13].

C. Transformers

Transformer has achieved multiple SOTA results in natural language processing tasks, and recently in computer vision tasks. Many works have explored the power of transformer in image captioning [22], [31]–[34]. For example, recently Marcella *et al.* [22] proposed a mesh-connected transformer framework for image captioning, making good use of each encoder layer's output as the input of each decoder layer.

The work in [32] introduced an entangled attention module which allows the transformer to exploit semantic and visual information simultaneously. However, the exploitation of transformer for medical report generation is still at the beginning [35]. The work in [35] proposed a memory-driven transformer for Chest X-ray report generation, which introduces a relational memory module to restore the information from previous generation processes. Please note that all above image captioners rely on CNNs to extract visual features for transformer. Outside image captioning, a very recent work in [9] first showed that a pure transformer without CNN could achieve SOTA performance for image classification. However, pure transformer framework has not been well explored in vision-to-language tasks, especially for medical report generation.

D. Image-Text Matching

Measuring the visual-semantic similarity between a given image-text pair, image-text matching learns strongly correlated visual and textual features. These models project both image and text features into a joint metric space and a triplet loss is minimized to ensure paired image-text samples close and unpaired samples apart [36]–[39].

Recently, image-text was leveraged to improve image captioning as in [23]. This work used a pretrained image-text matching model [37] to learn an attention map which further guides the training of the image captioning task. It is noted

that in [23], the two tasks of image-text matching and image captioning were trained separately, and they were linked via attention map. In contrast, in our work, image-text matching and report generation share the same network model and are jointly trained, so that they could fully negotiate with each other in learning visual and textual features.

It is also noted that recently transformer was used in image-text matching [40]. That work utilised a unified transformer encoder, fed the image and text features simultaneously and side by side, and performed a binary classification to tell whether the input pair is matched. Please note that, the work in [40] only involved transformer encoders for image-text matching. The situation becomes more complicated in our case since we need to at the same time consider the report generation task that involves both the encoder and decoder of the transformer.

III. PROPOSED METHOD

Fig. 1 shows the overview of our proposed model. It follows the transformer's encoder-decoder architecture and consists of a vision transformer encoder and a text transformer decoder. The vision encoder splits an image into patches as the input and uses a pure transformer to extract visual features. Its output is fed into both the text decoder to produce a report and the multi-label classification (MLC) network for diagnostic prediction. The text decoder is built upon a memory-augmented transformer which applies a memory-driven layer normalization module [35] to explore and preserve the relation between the words in a report. The generated reports are evaluated by the report generation (RG) loss via a term-weighting scheme. In addition to the main RG loss and the auxiliary MLC loss, another auxiliary image-text matching (ITM) loss is jointly assessed for training. It evaluates whether a given image-report pair is a true pair, enforcing strongly correlated visual and textual features to be learned for this case. The ITM task shares the same transformer structure as the RG task, with a temporal-weighted matching loss carefully designed by scrutinising transformer's masked self-attention mechanism. Our model components are elaborated as follows.

A. Structure of the Proposed Model

1) Vision Transformer Encoder: Let's denote an input image by $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels and H and W stand for image height and width, respectively. \mathbf{I} is reshaped as a sequence of flattened 2D image patches collectively represented by $\mathbf{I}_p \in \mathbb{R}^{N_p \times P^2 C}$, where “ p ” is short for “patch”, $N_p = \frac{HW}{P}$ is the total number of patches and P is the length of the side of a patch. In other words, the input image is represented by a vector with the dimensions of $N_p P^2 C$. The input image can now be viewed as a sequence of N_p image patch tokens, while the size of each token equals $P^2 C$.

These image patches are further processed by a linear projection layer and their embedding can be collectively expressed as $\mathbf{E}_p = \mathbf{I}_p \mathbf{W}_p^\top$, where $\mathbf{W}_p \in \mathbb{R}^{D \times P^2 C}$ and therefore \mathbf{E}_p is in the size of $N_p \times D$. Furthermore, at the head of patch

embedding, we concatenate a special $[CLS]$ token \mathbf{E}_{cls} like as in Bert [4]. It is a learnable one-dimensional embedding of the length D and its output from the transformer encoder (shown as the red block in Fig. 1) serves as the representation of the entire input image for classification. The concatenated patch embedding $[\mathbf{E}_{cls}; \mathbf{E}_p]$ is then added with a standard learnable 1D position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N_p+1) \times D}$ that carries the positional information, forming the input \mathbf{V}_0 to the transformer encoder:

$$\mathbf{Z}_0 = [\mathbf{E}_{cls}; \mathbf{E}_p] + \mathbf{E}_{pos}, \quad (1)$$

where “ $;$ ” indicates concatenation. \mathbf{Z}_0 consists of $(N_p + 1)$ row vectors $\{\mathbf{z}_0^i\}_{i=1}^{N_p+1}$, each of which represents the feature vector from an image patch.

As shown in Fig. 1, our vision transformer [8] encoder consists of three encoder layers (as indicated by “ $3 \times$ ” to the left of Fig. 1) and each layer contains several key components: multi-head self-attention (MSA), Feed Forward (FF), and Layer normalization (LN). The fundamental operation of MSA relies on the concept of self-attention, which is built upon three vectors: queries \mathbf{q} , keys \mathbf{k} and values \mathbf{v} . They are obtained by mapping the input \mathbf{z}_0^i into a metric space through a linear projection: $\mathbf{q}_i = \mathbf{W}_q \mathbf{z}_0^i$, $\mathbf{k}_i = \mathbf{W}_k \mathbf{z}_0^i$, and $\mathbf{v}_i = \mathbf{W}_v \mathbf{z}_0^i$, where matrices \mathbf{W}_q , \mathbf{W}_k , and \mathbf{W}_v are learnable parameters. The output of the MSA operation on \mathbf{z}_0^i is $\mathbf{a}_i = \text{softmax} \left(\frac{\mathbf{q}_i^\top \mathbf{k}}{\sqrt{d_k}} \right) \mathbf{v}_i$, where $i \in [1, N_p + 1]$ and $\mathbf{k} = [\mathbf{k}_1, \dots, \mathbf{k}_{N_p+1}]$, which can be considered as a weighted sum of the value vector while the weight is computed using a softmax activation function over the inner product of the query with the corresponding keys. All the computations can be performed in parallel and expressed as

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}} \right) \mathbf{V}. \quad (2)$$

The output of our transformer encoder is as follows:

$$\begin{aligned} \hat{\mathbf{Z}}_l &= \text{MSA}(\text{LN}(\mathbf{Z}_{l-1}) + \mathbf{Z}_{l-1}), l = 1, \dots, L_e \\ \mathbf{Z}_l &= \text{FF}(\text{LN}(\hat{\mathbf{Z}}_{l-1})) + \hat{\mathbf{Z}}_{l-1}, l = 1, \dots, L_e \end{aligned} \quad (3)$$

where L_e (i.e., 3 in our model) is the total number of encoder layers. $\hat{\mathbf{Z}}_l$ and \mathbf{Z}_l represent the outputs from the MSA and FF blocks in the l -th layer of the transformer encoder. The last output of the transformer encoder \mathbf{Z}_{L_e} contains two parts: the $[CLS]$ token embedding $\mathbf{Z}_{L_e}^{CLS}$ and the image patch embedding $\mathbf{Z}_{L_e}^{pat}$. The $\mathbf{Z}_{L_e}^{CLS}$ is fed into a multi-label classification task to predict the medical tags of the input image while the $\mathbf{Z}_{L_e}^{pat}$ is fed into the transformer decoder for report generation.

2) Text Transformer Decoder: The input of our text transformer decoder includes the vision encoder's output $\mathbf{Z}_{L_e}^{pat}$ and the right-shifted text sequence embedding \mathbf{T}_0 , which is the sum of the word embedding and the positional embedding, as done in common practice. As shown in Fig. 1, in addition to the common MSA and FF, we employ a memory-driven conditional layer normalization (MLN) [35]. Compared with the traditional layer normalization, MLN incorporates the relational memory module proposed in [35] by feeding its output m to γ and β . Consequently, this design can take the benefit

from the memory while preventing it from influencing too many parameters of Transformer so that some core information for generation is not affected. The expression of the memory-driven conditional layer normalization is shown as the follows.

$$\text{MLN} = \hat{\gamma} \odot \frac{r - \mu}{\sigma} + \hat{\beta} \quad (4)$$

where $\hat{\gamma}$ and $\hat{\beta}$ are expressed as $\hat{\gamma} = \gamma + f_{mlp}(m)$, $\hat{\beta} = \beta + f_{mlp}(m)$, respectively, in which $f_{mlp}(\cdot)$ is a linear projection function.

Our text decoder consists of three decoder layers $L_d = 3$, and the outputs of all previous layers are reused as the input in the successive layers. The outputs $\hat{\mathbf{T}}_l$, $\hat{\mathbf{T}}\mathbf{Z}_l$ and $\mathbf{T}\mathbf{Z}_l$ of the three blocks in each decoder layer are

$$\begin{aligned} \hat{\mathbf{T}}_l &= \text{MSA}(\text{MLN}(\mathbf{T}_{l-1})) + \mathbf{T}_{l-1} \\ \hat{\mathbf{T}}\mathbf{Z}_l &= \text{MSA}(\text{MLN}(\mathbf{Z}_{L_e}^{pat}, \mathbf{Z}_{L_e}^{pat}, \hat{\mathbf{T}}_l)) + \hat{\mathbf{T}}_l \\ \mathbf{T}\mathbf{Z}_l &= \text{FF}(\text{MLN}(\hat{\mathbf{T}}\mathbf{Z}_l)) + \hat{\mathbf{T}}\mathbf{Z}_l \end{aligned} \quad (5)$$

where $l \in \{1, \dots, L_d\}$ indexes the decoder layer.

Instead of just utilising the last decoder layer's output, we compute the average of the outputs from all the three decoder layers as the fused features of the image and text. This caters for both the general information contained in the earlier decoder layers and the task specific information contained in the later decoder layers.

B. Multi-Criteria Objectives

1) Term-Weighted Report Generation Loss: In medical report generation, the vocabulary of the reports usually faces the long-tail issue, that is, the key discriminative words occur at a low frequency in the reports while the less important words could appear much more frequently. In this case, treating each word equally will hurt the capability of the model in generating important information in a report. Therefore, assigning term weight to each word becomes necessary. The well-established TF-IDF measure in information retrieval is a natural choice for this purpose.

Let's denote the medical report and the vocabulary as $\{\mathbf{T}_k\}_{k=1}^{N_K}$ and $\{\mathbf{t}_s\}_{s=1}^{N_S}$, respectively. \mathbf{T}_k is the k th report in the corpus, and \mathbf{t}_s is the s th word in the vocabulary. Let n_{sk} be the total number of words \mathbf{t}_s in a report \mathbf{T}_k . Term Frequency and Inverse Document Frequency are defined as

$$\text{TF}_{sk} = \frac{n_{sk}}{\sum_{s=1}^{N_S} n_{sk}}; \quad \text{IDF}_s = \log \frac{N_K}{|\{k | \mathbf{t}_s \in \mathbf{T}_k\}|}. \quad (6)$$

The term weight of word \mathbf{t}_s in report \mathbf{T}_k is obtained as

$$tw_{sk} = \text{TF}_{sk} \cdot \text{IDF}_s. \quad (7)$$

We pre-compute the term weight of each word in a report as $\{tw_i\}_{i=1}^N$, where N denotes the total number of words in the report. By incorporating them into the cross-entropy loss commonly used in report generation, we produce a term-weighted report generation loss as follows.

$$\mathcal{L}_{RG}^{tw} = - \frac{\sum_{i=1}^N tw_i \cdot \log P(\mathbf{t}_i | \mathbf{I}, \mathbf{t}_{i-1}, \dots, \mathbf{t}_1)}{\sum_{i=1}^N tw_i}. \quad (8)$$

where $P(\mathbf{t}_i | \mathbf{I}, \mathbf{t}_{i-1}, \dots, \mathbf{t}_1)$ represents the probability predicted by the model for the i -th word based on the information of the image \mathbf{I} and the first $(i-1)$ words.

2) Temporal-Weighted ITM Loss: Some visual language tasks such as image-text matching explicitly learn image-text alignment. Compared with image captioning, those tasks do not have to consider sentence grammar and have more objective training losses [23]. In this paper, we leverage the joint training of image-text matching and report generation to learn strongly-correlated visual and textual features to facilitate report generation.

Image-text matching has to be tactically incorporated into our pure transformer-based report generation framework. As discussed in Section II, the sole image-text matching task could be easily embedded into transformer structure as this only touches the transformer's encoder and both the visual features and the text embedding are fed into the encoders concurrently. However, the situation becomes more complicated in our case as the image-text matching task is now entangled with the report generation task which contains both the encoder and decoder of the transformer. In this case, we propose to use the encoder for visual feature extraction and incorporate text embedding through the decoder. However, such incorporation is not straightforward since the report generation task and the image-text matching task share the same network model but have different ways of using the image and text features. For report generation, the transformer decoder takes the visual features and the embedding of the previous word as the input to generate the next word. This is implemented by covering the word embedding of the input sentence with a shift-right mask in order to generate the predicted word step by step and meanwhile, compute in parallel, which is known as masked self-attention (Fig. 2 (a)). However, this shift-right mask will cause an imbalance issue between image and text information during their feature fusion for image-text matching, since in the early steps of the masked self-attention (corresponding to the bottom rows of the mask matrix) a large portion of the sentence is masked off so that only a small portion of text information is used to match the whole image. The text information will gradually increase when more and more words are unmasked (corresponding to the top rows of the mask matrix) and finally the whole sentence is taken into account.

Scrutinising the masked self-attention mechanism, we design a temporal-weighted loss to compensate the imbalance of image and text information. The weighting function is a monotonically increasing one, giving higher weight to the later fused feature since it contains more textual content. It is formally expressed as follows. Let TE and TD be the embedding functions of the transformer's encoder and decoder. Let $\mathbf{T}_0 = \{\mathbf{t}_i\}_{i=1}^{N_r}$ be the shifted right input sequence of a report, where N_r is the sequence length. The decoder output is denoted as $\mathbf{F} = \text{TD}(\text{TE}(\mathbf{Z}_0), \mathbf{T}_0)$, where $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^{N_r}$ can be considered as the shift-righted fusion features. Specifically, \mathbf{f}_1 is the fusion feature coming from the visual feature \mathbf{Z}_0 and the first word \mathbf{t}_1 , while \mathbf{f}_{N_r} is the fusion feature from \mathbf{Z}_0 and the whole input sentence \mathbf{T}_0 .

A linear projection layer $\mathbf{S} = \mathbf{W}_f \mathbf{F}$ is further applied to map the fusion feature $\mathbf{F} \in \mathbb{R}^{N_r \times d_k}$ from d_k dimensions to

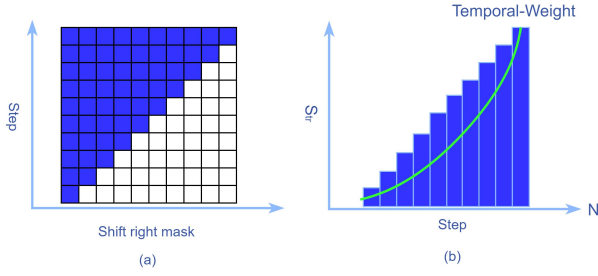


Fig. 2. (a) The shift-right mask: blue indicates “1” and white indicates “0”. (b) x-axis is the masking step, while y-axis is the ratio of text information (i.e., S_{tr}) in the similarity score. Due to the shift-right masked self-attention, this ratio increases with steps. Accordingly, we assign higher weight (green curve) to the similarity score of later steps to compute image-text matching loss.

one dimension. We interpret $\mathbf{S} = \{s_i\}_{i=1}^{N_r}$ as the similarity scores of the visual feature \mathbf{Z}_0 with the shift-right word tokens. As shown in Fig. 2 (b), with the increase of the step i , higher weight is assigned to the similarity score s_i . The overall similarity score $S(\mathbf{I}, \mathbf{T})$ between the input image \mathbf{I} and the input text sequence \mathbf{T} is defined as

$$S(\mathbf{I}, \mathbf{T}) = \frac{\sum_{i=1}^{N_r} \lambda_{ITM}^i \cdot s_i}{\sum_{i=1}^{N_r} \lambda_{ITM}^i}, \quad (9)$$

where $\lambda_{ITM}^i = (N_r - i + 1)^{-\alpha}$ increases temporally with step i . α is a preset hyper-parameter (e.g., 0.5 in this paper).

We train the image-text matching task with a triplet loss:

$$\mathcal{L}_{ITM} = [\max(0, S(\mathbf{I}, \mathbf{T}_{pos}) - S(\mathbf{I}, \mathbf{T}_{neg}) + \epsilon)] + \mu_{ITM} \cdot [S(\mathbf{I}, \mathbf{T}_{neg})^2 + S(\mathbf{I}, \mathbf{T}_{pos})^2], \quad (10)$$

where \mathbf{T}_{pos} (\mathbf{T}_{neg}) is the positive (negative) text sample that matches (mismatches) the given image \mathbf{I} . The negative samples are obtained by pairing each image with the report of another image from a min-batch. ϵ is a parameter of margin and μ_{ITM} is a hyper-parameter for L_2 -regularization.

3) Multi-Label Classification Loss: Upon the output of the transformer encoder, we introduce a multi-label classification (MLC) task to predict the medical tags of the input image to regularise report generation. Specifically, we only utilise the embedding of the $[CLS]$ token as the global representation of the image for this task, denoted as $V_L^{cls} \in \mathbb{R}^{b \times D}$, where b is the number of images in a mini-batch and D is the dimensions of the $[CLS]$ embedding. A linear projection maps V_L^{cls} from D to K dimensions, where K is the number of multi-label classes. The MLC loss is defined as

$$\mathcal{L}_{MLC} = -\frac{1}{K} \cdot \sum_i y_i \cdot \log((1 + \exp(-x_i))^{-1}) + (1 - y_i) \cdot \log\left(\frac{\exp(-x_i)}{1 + \exp(-x_i)}\right), \quad (11)$$

where x_i is the prediction for tag i ($i \in \{0, 1, \dots, K\}$), and y_i is the ground-truth label for tag i where $y_i \in \{0, 1\}$, with $y_i = 1$ meaning the input image has the corresponding tag while $y_i = 0$ meaning the opposite.

4) Overall Objective Function: Our overall objective integrates the three losses regarding report generation, image-text matching, and classification, which is defined as

$$\mathcal{L}_{all} = \lambda_{RG} \mathcal{L}_{RG}^{tw} + \lambda_{ITM} \mathcal{L}_{ITM} + \lambda_{MLC} \mathcal{L}_{MLC}. \quad (12)$$

The hyper-parameters λ_{RG} , λ_{ITM} and λ_{MLC} balance the three loss terms, and their values are given in Section. IV.

IV. EXPERIMENTS

A. Datasets

In this paper, we evaluate the performance of our proposed model and compare it with the state-of-the-arts (SOTA) image captioning and medical report generation methods on two public datasets.

1) IU-Xray: As the most widely used benchmark for medical report generation task, Indiana University Chest X-ray Collection (IU-Xray) [43] contains 3,955 fully de-identified radiology reports, each with a frontal and/or lateral chest X-ray images, counting to 7,470 chest X-ray images in total. Each report is composed of sections such as “Impression”, “Findings” and “Tags”, where “Impression” is a single sentence description of the image, “Findings” is a long paragraph with detailed description of the evidence and “Tags” are the keywords which are identified using the Medical Text Indexer.¹ In this work, by considering only the reports with two complete image views and the complete sections of “Findings” and “Impression”, we work on a smaller dataset with 3195 reports and 6390 images. All words in “Findings” and “Impression” are tokenised into 2,076 unique words, with two special tokens, $\langle start \rangle$ and $\langle end \rangle$, indicating the start and the end of a sentence. Following [12], [17], 10% of reports are randomly picked to form the test set. All evaluations are done on the test set.

2) MIMIC-CXR: The recently released MIMIC-CXR [44] is the largest public dataset containing both the chest radiographs and free-text reports. In total, it consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016 [45]. In our experiment, we adopt MIMIC-CXR’s official split following the work [35] for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test, respectively. We convert all tokens to lowercase characters and remove non-word illegal characters and the words whose occurring frequency lower than 10, counting to 4030 unique words remaining in the dataset. In addition, the structured labels provided by this dataset, including “No Finding” and other 13 descriptive labels, are used as the 14 tags to train our MLC task. Each label contains one of the four values: 1, 0, -1, or “missing”, representing “positively mentioned”, “negatively mentioned”, “mentioned with uncertainty” and “no mention”, respectively. We only keep the labels with the values 1 and 0 when training the MLC task.

¹<https://ii.nlm.nih.gov/MTI/>

TABLE I

COMPARISON ON IU-XRAY (UPPER PART) AND MIMIC-CXR DATASETS (LOWER PART). † AND †† INDICATES THE RESULTS ARE QUOTED FROM THE PUBLISHED LITERATURE. SPECIFICALLY, WE QUOTE THE RESULTS FROM [41] FOR IU-XRAY, AND FROM [35] FOR MIMIC-CXR. FOR OTHER METHODS IN COMPARISON, THEIR RESULTS ARE OBTAINED BY RE-RUNNING THE PUBLICLY RELEASED CODES ON THESE TWO DATASETS USING THE SAME TRAINING-TEST PARTITION AS OUR METHOD SO THAT THEY ARE STRICTLY COMPARABLE. PLEASE NOTE THAT THE QUOTED RESULTS INDICATED BY †† MAY NOT BE STRICTLY COMPARABLE SINCE THE DATA SPLIT WAS UNKNOWN FOR THEM

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
IU-Xray	Show-Tell [1]	0.346	0.214	0.141	0.095	0.320	0.239
	Att2in [28]	0.399	0.249	0.172	0.126	0.321	0.341
	AdaAtt [7]	0.436	0.288	0.203	0.150	0.354	0.265
	Topdown [21]	0.375	0.237	0.167	0.123	0.319	0.336
	Transformer [8]	0.422	0.264	0.177	0.120	0.338	0.268
	M2transformer [22]	0.463	0.318	0.214	0.155	0.335	0.349
	Grounded [23]	0.446	0.301	0.237	0.176	0.343	0.395
	MRMA [17]	0.415	0.315	0.223	0.166	0.349	0.176
	R2Gen [35]	0.445	0.283	0.202	0.155	0.372	0.353
	CoAtt† [12]	0.455	0.288	0.205	0.154	0.369	0.277
	HGRG-Agent† [42]	0.438	0.298	0.208	0.151	0.322	0.343
	CMAS-RL† [41]	0.464	0.301	0.210	0.154	0.362	0.275
	Ours	0.496	0.319	0.241	0.175	0.377	0.449
MIMIC-CXR	Show-Tell† [1]	0.299	0.184	0.121	0.084	0.263	-
	Att2in† [28]	0.325	0.203	0.136	0.096	0.276	-
	AdaAtt† [7]	0.299	0.185	0.124	0.088	0.266	-
	Transformer† [8]	0.314	0.192	0.127	0.090	0.265	-
	M2transformer [22]	0.212	0.128	0.083	0.058	0.240	0.074
	Grounded [23]	0.271	0.174	0.122	0.094	0.257	0.150
	R2Gen [35]	0.347	0.212	0.143	0.103	0.271	0.142
	Ours	0.351	0.223	0.157	0.118	0.287	0.281

B. Experimental Settings

1) *Evaluation Metrics*: Following the standard evaluation protocol,² we utilise the most widely used BLEU scores [46], ROUGE-L [47] and CIDER [48] as the metrics to evaluate the quality of the generated text report.

2) *Implementation Details*: We use paired images as the input for IU-Xray and concatenate the features of the frontal and lateral views, while use single image input for MIMIC-CXR to ensure consistency with the experiment settings in [35]. We also adopted the same preprocessing to the reports as paper [35], removed irregular characters and uniformly replaced the connectors between sentences in the report with “. ”. For vision transformer encoder, we resize the input image to 448×448 and set the patch size to 32. The number of the layers in both the vision transformer encoder and the text transformer decoder is set to 3 and the number of heads in multi-head attention is set to 8. The hyper-parameters ϵ and μ_{ITM}^{Reg} in Eqn. 10 are set to 0.05 and 0.025, while λ_{RG} , λ_{MLC} , and λ_{ITM} in Eqn. 12 are set as 1, 2, and 5, respectively. We train our model using Adam optimizer [49] with mini-batch size of 16. The learning rates are set to be $5e-5$ and $1e-4$, respectively, for the vision encoder and the rest parameters. The model is trained in a total of 30 epochs. These settings are equally applied to both datasets. For the model’s efficiency, in the training phase, we use two NVIDIA 2080Ti GPUs and trained our model for about 48 hours on the MIMIC-CXR dataset and 3 hours on the IU-Xray dataset in a total of 30 epochs. In the inference phase, the inference time per batch is about 4.5 seconds, with a batch size of 32 and a maximum report length of 60.

C. Results and Discussion

1) *Comparison With SOTA Methods*: Table I shows the performance comparison between our proposed model and existing SOTA methods in image captioning and medical report generation.

On IU-Xray, there are seven SOTA image captioning methods in the comparison, including Show-tell [1], AdaAtt [7], Att2in [21], Topdown [28], Transformer [8], M2transformer [22], and Grounded [23]. The results of these methods are obtained based on our re-running of the publicly released codes under the same experiment setting as ours, so they are strictly comparable. Moreover, we also compare with five SOTA medical report generation methods: including CoAtt [12], HGRN-Agent [42], KERP [41], MRMA [17], and the very recent method R2Gen [35] which incorporated transformer (as the text decoder) for Chest X-ray report generation. For MRMA [17] and R2Gen [35], we download the publicly released code and re-run it using the same experiment setting as ours. For the rest three methods CoAtt [12], HGRN-Agent [42] and KERP [41], we have to directly quote their results from [35], [41] since we could not access their codes. These results were chosen because the performance of the basic Show-Tell method reported in [35], [41] is close to that obtained in our experiment, implying certain comparison base. However, since the data split is unknown, the quoted results of CoAtt [12], HGRN-Agent [42] and KERP [41] may not be strictly comparable with that of our method, and they were listed just for reference. As shown in Table I (upper part), on IU-Xray, our proposed method achieves the best performance over almost all evaluation metrics among the comparing methods. It outperforms the second best performer Grounded [23] that also makes use of image-text match-

²<https://github.com/tylin/coco-caption>

TABLE II

ABLATION STUDY ON KEY COMPONENTS USING IU-XRAY. HERE “BASELINE” REFERS TO OUR PURE TRANSFORMER FRAMEWORK WITHOUT THE THREE CRITERIA. “MLC”, “ITM”, “TW”, AND “TMW” STAND FOR “MULTI-LABEL CLASSIFICATION”, “IMAGE-TEXT MATCHING”, “TERM WEIGHTING”, AND “TEMPORAL-WEIGHTING”. B@4 IS THE BLUE SCORE COMPUTED WITH 4-GRAMS

Model	B@4	Rouge	CIDEr
baseline	0.151	0.367	0.355
baseline+MLC	0.162	0.372	0.381
baseline+ITM+MLC	0.169	0.375	0.421
Ours\TmW	0.157	0.354	0.399
Ours\ITM	0.164	0.371	0.407
Ours (baseline+ITM+MLC+TW)	0.175	0.377	0.449

ing for image captioning over all metrics except BLEU-4. This demonstrates the advantages of our joint training of image-text matching and report generation over the separated manipulation of the two tasks in [23] (discussed in Section II). Meanwhile, it is noted that our proposed model also beats the three transformer based models Transformer [8], M2transformer [22], and R2Gen [35] in comparison.

On MIMIC-CXR, we compare with R2Gen [35] and another six image captioners as shown in Table I (lower part). Except M2transformer [22], Grounded [23] and R2Gen [35], the results of other comparing methods are quoted from [35]. Please note that since the official split of MIMIC-CXR is used in all these methods and ours, the results are comparable. Again, our model is overall the best performer on this largest benchmark. It wins the vanilla Transformer [4], M2transformer [22], and R2Gen [35], all of which still need CNNs in addition to transformer.

From the results on the two benchmarks, we observe that our proposed model could especially improve CIDEr that down-weights common words in all reports. This reflects the effectiveness of our strategies in tightly aligning visual and textual features and weighting the importance of words.

2) Ablation Study: Based on IU-Xray, We conduct an ablation study on the key components of our proposed model, singling out the contributions of each component. For this purpose, we remove the three criteria from our proposed method as the baseline, which is a pure transformer model (with the memory-augmented module) supervised by the common cross-entropy loss for report generation, to verify the performance improvements brought by Temporal-weighted ITM supervision, MLC supervision and the term-weighted RG supervision. In addition, to further investigate the ITM loss with our proposed temporal-weighting, we also test two variants: one completely removes the ITM loss (together with the temporal weighting) and is denoted as “Ours_{setminus}ITM”; the other keeps the ITM loss but switches off the temporal weighting in our proposed model by setting all temporal weights to 1, and is denoted as “Ours\TmW”.

As can be seen, using our supervision criteria, our model could significantly elevate the performance of our baseline transformer model in all three metrics. The most remarkable overall improvement is in CIDEr, which increases from 0.355 to 0.449. This is a favorable property since CIDEr focuses more on those words that discriminate across reports.

TABLE III

ABLATION STUDIES ON HYPER-PARAMETERS USING IU-XRAY. WE FIX $\lambda_{RG} = 1$ AND VARY λ_{ITM} AND λ_{MLC} FOR THE INVESTIGATION. THE DEFAULT VALUES ARE: $\lambda_{RG} = 1$, $\lambda_{ITM} = 5$, AND $\lambda_{MLC} = 2$

λ_{ITM}	λ_{MLC}	B@4	Rouge	CIDEr
1	1	0.165	0.367	0.401
5	2	0.175	0.377	0.449
10	2	0.157	0.302	0.374
5	4	0.171	0.373	0.422

TABLE IV

ABLATION STUDIES ON HYPER-PARAMETERS OF EQN. 9 USING IU-XRAY

ϵ	μ_{ITM}	B@4	Rouge	CIDEr
0.05	0.025	0.175	0.377	0.449
0.025(0.05/2)	0.025	0.173	0.374	0.437
0.1(0.05*2)	0.025	0.177	0.378	0.452
0.05	0.0125(0.025/2)	0.170	0.372	0.413
0.05	0.05(0.025*2)	0.174	0.375	0.434

Specifically, the regularisation from MLC loss could improve the baseline in all three metrics, with BLUE-4 from 0.151 to 0.162, Rouge from 0.367 to 0.372, and CIDEr from 0.355 to 0.381. This performance could be further increased by the addition of the ITM loss, with BLUE-4 from 0.162 to 0.169, Rouge from 0.372 to 0.375, and more significantly CIDEr from 0.381 to 0.421. Eventually, the best performance is achieved by our proposed whole model (denoted as “Ours”) that further employs term weighting to raise the importance of key words. Furthermore, comparing “Ours” and “Ours\TmW”, it can be seen that our proposed temporal weighting in ITM loss could significantly improves BLUE4 from 0.157 to 0.175, and CIDEr from 0.399 to 0.449, verifying our analysis in ITM supervision. The importance of the temporal weighting to the ITM loss could also be reflected by comparing the results of “Ours\ITM” and “Ours\TmW”. It implies that without the proposed temporal weighting, the addition of ITM loss may negatively affect the performance, which is consistent with our observation that the report generation task and the ITM task have different ways of using the fused image and text features and thus integrating them in the same network structure has to be carefully designed like in our approach.

we further perform an ablation study to analyze the hyper-parameters in Eqn. 9. The Eqn. 9 contains two hyper-parameters: ϵ and μ_{ITM} , where ϵ is the margin of triplet loss and μ_{ITM} is a hyper-parameter for L2-regularization. We fix one hyper-parameter and increase/decrease the other parameter by 2 times, respectively. As shown in Table IV, we can achieve a better result when we set ϵ and μ_{ITM} as 0.1 and 0.025 respectively.

Finally, we perform another ablation study based on IU-Xray to investigate the influence of key hyper-parameters λ_{RG} , λ_{ITM} , and λ_{MLC} used to balance our proposed three supervision criteria. We fix $\lambda_{RG} = 1$, and vary the other two parameters as in Table III for investigation. As can be seen, when these hyper-parameters vary in a reasonably large range, our model can still perform better than most of the methods in comparison in Table I (upper part).

3) Qualitative Results: Fig. 3 and Fig. 4 exhibit the generated reports on the IU-Xray dataset and MIMIC-CXR dataset, respectively. To facilitate comparison, we highlight the key

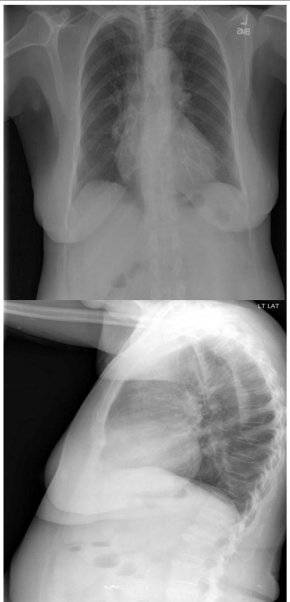
	Ground-truth	M2transformer	Grounded
	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly. Normal pulmonary vascularity. Tortuosity of the descending aorta. No focal infiltrate, pneumothorax or pleural effusion.	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly. the lungs are clear of focal airspace disease pneumothorax or pleural effusion . No focal consolidation, pneumothorax, or effusion. Relative elevation of the left hemidiaphragm noted. there are no acute bony findings	Impression: Cardiomegaly with elevated right hemidiaphragms. Findings: Mild cardiomegaly. Low lung volumes with elevated hemidiaphragms, greater on the right, there is no focal air space opacity . no pleural effusion or pneumothorax . the osseous structures are intact
	R2Gen	Baseline	Ours
	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly. Normal pulmonary vascularity. No focal infiltrate or consolidation, there is no focal air space opacity , Pulmonary vascularity within normal limits. no acute bony abnormality	Impression: no acute cardiopulmonary findings . Findings: the cardio mediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly. Normal pulmonary vascularity. No focal infiltrate, pneumothorax or pleural effusion, there is no focal air space opacity , no acute bony abnormality

Fig. 3. Examples of the generated reports on IU-xray dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer [22], Grounded [23] and R2Gen [35]. The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth.


	Ground-truth	M2transformer	Grounded
	Impression: No acute cardiopulmonary process Findings: Frontal and lateral views of the chest. Improved inspiratory effort seen on the current exam. The lungs are clear of consolidation, effusion, or pulmonary vascular congestion. The cardio mediastinal silhouette is within normal limits. No acute osseous abnormalities identified.	Impression: no acute cardiopulmonary process . Findings: the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable.	Impression: No acute cardiopulmonary process. Findings: The lungs are clear. The cardio mediastinal silhouette is within normal limits. No acute osseous abnormalities.
	R2Gen	Baseline	Ours
	Impression: No acute cardiopulmonary process Findings: Frontal and lateral radiographs of the chest were acquired. The lungs are clear of consolidation or pulmonary vascular congestion. Cardio mediastinal silhouette is unchanged noting unfolding of the descending thoracic aorta. No acute osseous abnormality detected.	Impression: No acute cardiopulmonary process. Findings: Frontal and lateral views of the chest. There is no pleural effusion, pneumothorax or focal airspace consolidation. Elevation of the left hemidiaphragm is unchanged. The heart size is normal and the mediastinal contours are unremarkable.	Impression: No acute cardiopulmonary process. Findings: Frontal and lateral views of the chest are obtained. The lungs are clear of focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. There is no pulmonary edema. No acute osseous abnormalities identified.

Fig. 4. Examples of the generated reports on MIMIC-CXR dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer [22], Grounded [23] and R2Gen [35]. The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth.

information in ground truth reports in bold. Meanwhile, in the generated reports, sentences are highlighted with different colors: red for incorrect generations (content not existing in the ground-truth reports), blue for partially correct generations (similar meaning but with different expressions), and green for completely correct generations (same expressions). As can be observed, compared with the baseline and the other three competitive methods, the report generated by our full

model captures more key information under the supervision of the three criteria, producing more sentences in green, which indicates our generated reports are more consistent with the ground-truth reports.

We also add two failure cases as shown in Fig. 5 for IU-Xray (upper) and MIMIC-CXR (lower) datasets. After analyzing the failure cases, we think one important reason causing the failures might be the long-tailed distribution of sentences/terms

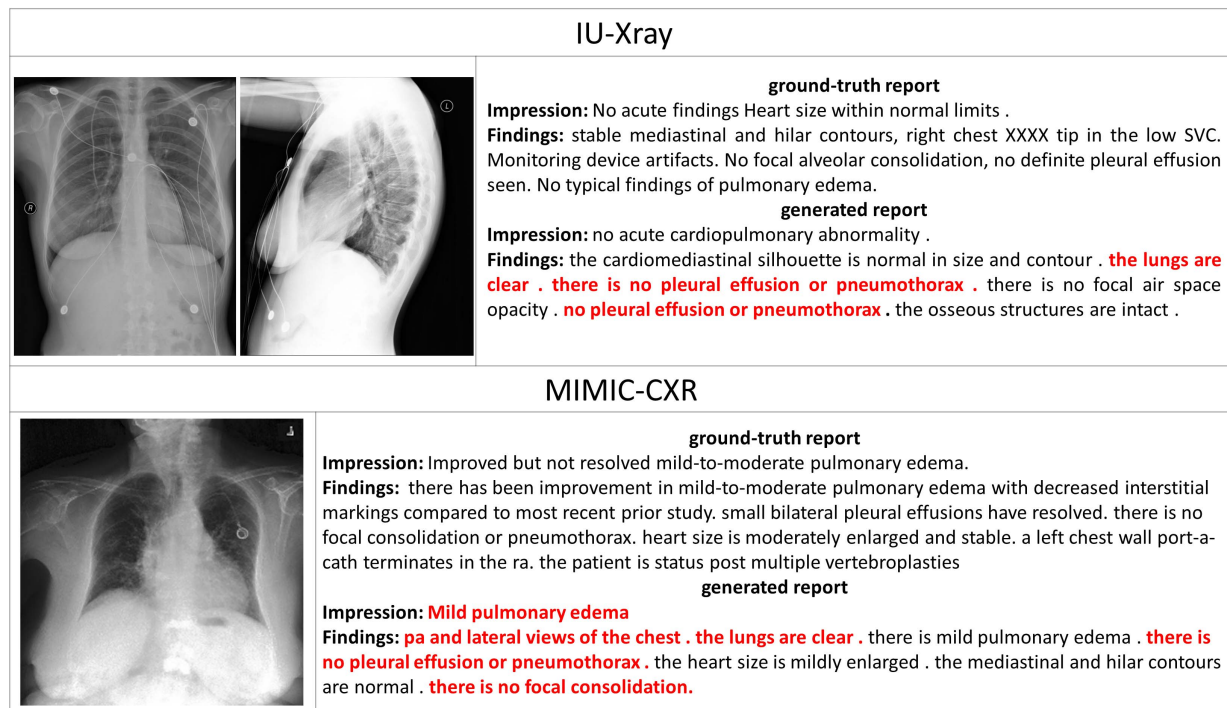


Fig. 5. The failure cases for IU-xray dataset (upper case) and MIMIC-CXR dataset (lower case). Sentences marked in red indicate frequently occurring sentences in the report.

in the reports of these datasets. The most commonly used templates in IU-Xray are counted in paper [50], including: “The lungs are clear”, “There is no pleural effusion or pneumothorax.”, “No evidence of focal consolidation, pneumothorax, or pleural effusion”, etc. Just like the failure case shown in Fig. 5, those sentences (marked in red) are more likely to be predicted by the model since they appear frequently in the report. Furthermore, for the MIMIC-CXR dataset, we did our own statistics from 222758 training reports to analyze the case shown in Fig. 5 (lower part). For the generated report, the sentence “pa and lateral views of the chest” appeared 22678 times in the training set, “the lungs are clear” appear 40038 times, the sequence “mild pulmonary edema” appear 5109 times; For the ground-truth report, the term “mild-to-moderate pulmonary edema” appears 432 times and the sentence “small bilateral pleural effusions have resolved” appears only 19 times in all training reports, while they are missing in the generated report. From the above analysis, we can see that the problem of long-tail distribution does exist and this will be a very valuable research direction in the future.

V. CONCLUSION

This work is the first attempt to build pure transformer based model to accurately generate diagnostic reports from radiographs. The proposed model is effectively trained under three supervision criteria to ensure well aligned visual and textual features, accurate multi-label diagnostic discrimination, and word-importance aware report generation. In this way, the generated reports are sensitive to similar images and key words of clinic importance, offering the state-of-the-arts performance on two benchmarks. Although our method achieved promising

results in medical report generation, there are some limitations in the current model. First, the example failure cases we provided in Fig. 5 show the evidence about the problem of long-tail distribution of terms in the training reports. There are some efforts exploring long-tail distribution in the field of natural image classification and object detection, but such research in the natural language processing and medical report generation is still superficial and could be taken as a future research direction. Second, we apply global self-attention for the report, making the calculation relatively expensive. Efficient transformers [51] could be explored in the future for medical report generation. Furthermore, as for our future work, the medical text indexer can be utilized to obtain more fine-grained tags for multi-label classification task. Moreover, knowledge graphs could potentially be built into our transformer-based framework for further improvement with additional information, as knowledge graphs may be beneficial to the generation of medical reports due to their ability to identify and model the relationships between different medical terms. However, the knowledge graphs constructed in the existing works [14], [52] are pretty sparse with very minimal information, requiring further improvement in future.

REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.
- [2] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Assoc. Comput. Linguistics*, 2019.
- [5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
- [6] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov, "Review networks for caption generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.
- [7] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.
- [8] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.
- [9] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.
- [10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.
- [11] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Trans. Multimedia*, vol. 23, pp. 2413–2427, 2021.
- [12] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," in *Proc. Assoc. Comput. Linguistics*, 2018.
- [13] C. Yin *et al.*, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 728–737.
- [14] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12910–12917.
- [15] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 317–325.
- [16] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 721–729.
- [17] Y. Xue *et al.*, "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 457–466.
- [18] Y. Xue and X. Huang, "Improved disease classification in chest X-rays with transferred features from report generation," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2019, pp. 125–138.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [21] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [22] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
- [23] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4777–4786.
- [24] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 684–699.
- [25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [26] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
- [27] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
- [28] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.
- [29] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 873–881.
- [30] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10685–10694.
- [31] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2020.
- [32] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8928–8937.
- [33] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 153–169.
- [34] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2019.
- [35] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020.
- [36] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf.*, 2018.
- [37] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
- [38] C. Liu, Z. Mao, A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 3–11.
- [39] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–34.
- [40] Y. C. Chen *et al.*, "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.
- [41] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6666–6673.
- [42] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2018.
- [43] D. F. Dina *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 2, no. 3, pp. 304–310, 2015.
- [44] A. E. W. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.
- [45] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, "Diagnostic captioning: A survey," 2021, *arXiv:2101.07299*.
- [46] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Assoc. Comput. Linguistics*, 2002, pp. 1–8.
- [47] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Assoc. Comput. Linguistics*, 2004, pp. 74–81.
- [48] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [50] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019.
- [51] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020, *arXiv:2009.06732*.
- [52] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13753–13762.