IIHT: Medical Report Generation with Image-to-Indicator Hierarchical Transformer

Keqiang Fan(✉), Xiaohao Cai, and Mahesan Niranjan

School of Electronics and Computer Science, University of Southampton,
Southampton SO17 1BJ, UK
{k.fan,x.cai,mn}@soton.ac.uk

Abstract. Automated medical report generation has become increasingly important in medical analysis. It can produce computer-aided diagnosis descriptions and thus significantly alleviate the doctors' work. Inspired by the huge success of neural machine translation and image captioning, various deep learning methods have been proposed for medical report generation. However, due to the inherent properties of medical data, including data imbalance and the length and correlation between report sequences, the generated reports by existing methods may exhibit linguistic fluency but lack adequate clinical accuracy. In this work, we propose an image-to-indicator hierarchical transformer (IIHT) framework for medical report generation. It consists of three modules, i.e., a classifier module, an indicator expansion module and a generator module. The classifier module first extracts image features from the input medical images and produces disease-related indicators with their corresponding states. The disease-related indicators are subsequently utilised as input for the indicator expansion module, incorporating the "data-text-data" strategy. The transformer-based generator then leverages these extracted features along with image features as auxiliary information to generate final reports. Furthermore, the proposed IIHT method is feasible for radiologists to modify disease indicators in real-world scenarios and integrate the operations into the indicator expansion module for fluent and accurate medical report generation. Extensive experiments and comparisons with state-of-the-art methods under various evaluation metrics demonstrate the great performance of the proposed method.

Keywords: Medical report generation · Deep neural networks · Transformers · Chest X-Ray

1 Introduction

Medical images (e.g. radiology and pathology images) and the corresponding reports serve as critical catalysts for disease diagnosis and treatment [22]. A medical report generally includes multiple sentences describing a patient's history symptoms and normal/abnormal findings from different regions within the medical images. However, in clinical practice, writing standard medical reports

B. Luo et al. (Eds.): ICONIP 2023, LNCS 14452, pp. 57–71, 2024.
https://doi.org/10.1007/978-981-99-8076-5_5

is tedious and time-consuming for experienced medical doctors and error-prone for inexperienced doctors. This is because the comprehensive analysis of e.g. X-Ray images necessitates a detailed interpretation of visible information, including the airway, lung, cardiovascular system and disability. Such interpretation requires the utilisation of foundational physiological knowledge alongside a profound understanding of the correlation with ancillary diagnostic findings, such as laboratory results, electrocardiograms and respiratory function tests. Therefore, the automatic report generation technology, which can alleviate the medics' workload and effectively notify inexperienced radiologists regarding the presence of abnormalities, has garnered dramatic interest in both artificial intelligence and clinical medicine.

Medical report generation has a close relationship with image captioning [9,31]. The encoder-decoder framework is quite popular in image captioning, e.g., a CNN-based image encoder to extract the visual information and an RNN/LSTM-based report decoder to generate the textual information with visual attention [11,14,29,30]. With the recent progress in natural language processing, investigating transformer-based models as alternative decoders has been a growing trend for report generation [2,3,21,27]. The self-attention mechanism employed inside the transformer can effectively eliminate information loss, thereby maximising the preservation of visual and textual information in the process of generating medical reports. Although these methods have achieved remarkable performance and can obtain language fluency reports, limited studies have been dedicated to comprehending the intrinsic medical and clinical problems. The first problem is *data imbalance*, e.g., the normal images dominate the dataset over the abnormal ones [24] and, for the abnormal images, normal regions could encompass a larger spatial extent than abnormal regions [17]. The narrow data distribution could make the descriptions of normal regions dominate the entire report. On the whole, imbalanced data may degrade the quality of the automatically generated reports, or even result in all generated reports being basically similar. The second problem is *length and correlation between report sequences*. Medical report generation is designed to describe and record the patient's symptoms from e.g. radiology images including cardiomegaly, lung opacity and fractures, etc. The description includes various disease-related symptoms and related topics rather than the prominent visual contents and related associations within the images, resulting in the correlation inside the report sequences not being as strong as initially presumed. The mere combination of encoders (e.g. CNNs) and decoders (e.g. RNN, LSTM, and transformers) is insufficient to effectively tackle the aforementioned issues in the context of medical images and reports since these modalities represent distinct data types. The above challenges motivate us to develop a more comprehensive method to balance visual and textual features in unbalanced data for medical report generation.

The radiologists' working pattern in medical report writing is shown in Fig. 1. Given a radiology image, radiologists first attempt to find the abnormal regions and evaluate the states for each disease indicator, such as uncertain, negative and

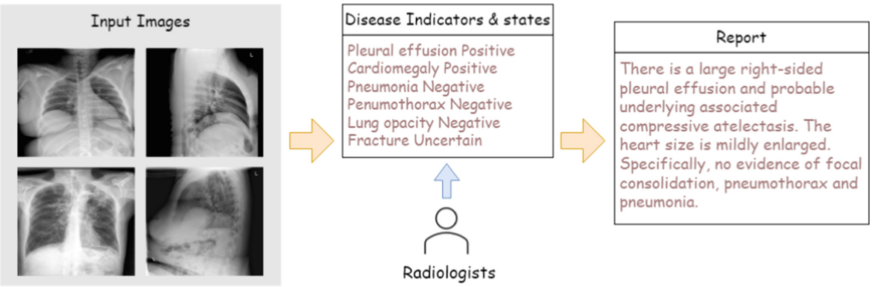

Input Images
Disease Indicators & states
Pleural effusion Positive
Cardiomegaly Positive
Pneumonia Negative
Penumothorax Negative
Lung opacity Negative
Fracture Uncertain
Radiologists
Report
There is a large right-sided pleural effusion and probable underlying associated compressive atelectasis. The heart size is mildly enlarged. Specifically, no evidence of focal consolidation, pneumothorax and pneumonia.

Fig. 1. The medical report writing procedure undertaken by radiologists.

positive. Then a correct clinical report is written through the stages for different indicators based on their working experience and prior medical knowledge. In this paper, we propose an image-to-indicator hierarchical transformer (IIHT) framework, imitating the radiologists' working patterns (see Fig. 1) to alleviate the above-mentioned problems in medical report generation.

Our IIHT framework models the above working patterns through three modules: *classifier*, *indicator expansion* and *generator*. The classifier module is an image diagnosis module, which could learn visual features and extract the corresponding disease indicator embedding from the input image. The indicator expansion module conducts the data-to-text progress, i.e., transferring the disease indicator embedding into short text sequences. The problem of data imbalance could be alleviated by encoding the indicator information, which models the domain-specific prior knowledge structure and summarises the disease indicator information and thus mitigates the long-sequence effects. Finally, the generator module produces the reports based on the encoded indicator information and image features. The whole generation pipeline is given in Fig. 2, which will be described in detail in Sect. 3. We remark that the disease indicator information here can also be modified by radiologists to standardise report fluency and accuracy. Overall, the contributions of this paper are three-fold:

- We propose the IIHT framework, aiming to alleviate the data bias/imbalance problem and enhance the information correlation in long report sequences for medical report generation.
- We develop a dynamic approach which leverages integrated indicator information and allows radiologists to further adjust the report fluency and accuracy.
- We conduct comprehensive experiments and comparisons with state-of-the-art methods on the IU X-Ray dataset and demonstrate that our proposed method can achieve more accurate radiology reports.

The rest of the paper is organised as follows. Section 2 briefly recalls the related work in medical report generation. Our proposed method is introduced in Sect. 3. Sections 4 and 5 present the details of the experimental setting and corresponding results, respectively. We conclude in Sect. 6.

2 Related Work

Image Captioning. The image captioning methods mainly adopt the encoder-decoder framework together with attention mechanisms [31] to translate the image into a single short descriptive sentence and have achieved great performance [1,15,18,26]. Specifically, the encoder network extracts the visual representation from the input images and the decoder network generates the corresponding descriptive sentences. The attention mechanism enhances the co-expression of the visual features derived from the intermediate layers of CNNs and the semantic features from captions [31]. Recently, inspired by the capacity of parallel training, transformers [25] have been successfully applied to predict words according to multi-head self-attention mechanisms. However, these models demonstrate comparatively inferior performance on medical datasets as opposed to natural image datasets, primarily due to the disparity between homogeneous objects observed in different domains. For instance, in the context of X-Ray images, there exists a relatively minimal discernible distinction between normal and abnormal instances, thereby contributing to the challenge encountered by models in accurately generating such captions.

Medical Report Generation. Similar to image captioning, most existing medical report generation methods attempt to adopt a CNN-LSTM-based model to automatically generate fluent reports [11,14,20,28]. Direct utilisation of caption models often leads to the generation of duplicate and irrelevant reports. The work in [11] developed a hierarchical LSTM model and a co-attention mechanism to extract the visual information and generate the corresponding descriptions. Najdenkoska et al. [20] explored variational topic inference to guide sentence generation by aligning image and language modalities in a latent space. A two-level LSTM structure was also applied with a graph convolution network based on the knowledge graph to mine and represent the associations among medical findings during report generation [27]. These methodologies encompass the selection of the most probable diseases or latent topic variables based on the sentence sequence or visual features within the data in order to facilitate sentence generation. Recently, inspired by the capacity of parallel training, transformers [13,32] have successfully been applied to predict words according to the extracted features from CNN. Chen et al. [2] proposed a transformer-based cross-modal memory network using a relational memory to facilitate interaction and generation across data modalities. Nguyen et al. [21] designed a differentiable end-to-end network to learn the disease feature representation and disease state to assist report generation.

The existing methods mentioned above prioritise the enhancement of feature alignment between visual regions and disease labels. However, due to the inherent data biases and scarcity in the medical field, these models exhibit a bias towards generating reports that are plausible yet lack explicit abnormal descriptions. Generating a radiology report is very challenging as it requires the contents of key medical findings and abnormalities with detailed descriptions for different data modalities. In this study, we address the challenges associated with data bias and scarcity in clinical reports through the utilisation of disease indicators as a bridge for more comprehensive medical report generation.

3 Method

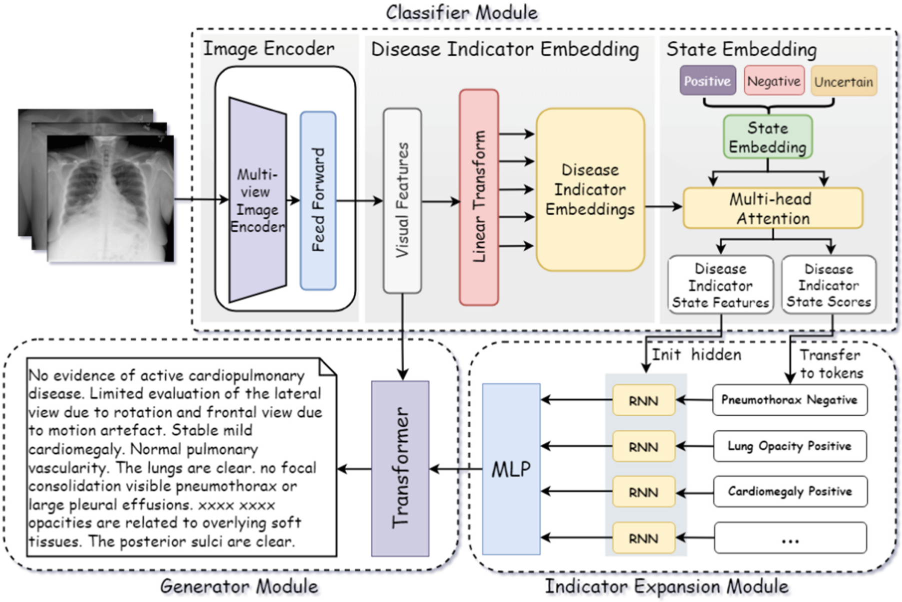

Classifier Module
Image Encoder
Multi-view Image Encoder
Feed Forward
Disease Indicator Embedding
Visual Features
Linear Transform
Disease Indicator Embeddings
State Embedding
Positive
Negative
Uncertain
State Embedding
Multi-head Attention
Disease Indicator State Features
Disease Indicator State Scores
Init hidden
Transfer to tokens
RNN
RNN
RNN
RNN
Pneumothorax Negative
Lung Opacity Positive
Cardiomegaly Positive
...
MLP
Transformer
No evidence of active cardiopulmonary disease. Limited evaluation of the lateral view due to rotation and frontal view due to motion artefact. Stable mild cardiomegaly. Normal pulmonary vascularity. The lungs are clear. no focal consolidation visible pneumothorax or large pleural effusions. xxxx xxxx opacities are related to overlying soft tissues. The posterior sulci are clear.
Generator Module
Indicator Expansion Module

Fig. 2. The proposed IIHT framework. It consists of three modules: classifier, indicator expansion and generator.

An overview of our proposed IIHT framework is demonstrated in Fig. 2. It follows the distinct stages involved in generating a comprehensive medical imaging diagnosis report, adhering to the established process employed in clinical radiology (e.g. see Fig. 1).

Given a radiology image $\mathbf{I}$, the corresponding different indicators are all classified into different states (e.g. positive, negative, uncertain, etc.) denoted as $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_t, \cdots, \mathbf{c}_T\}$, where T is the number of indicators and $\mathbf{c}_t$ is the one-hot encoding of the states. Particularly, these indicators can also be modified by radiologists to standardise the disease states across patients, thereby enhancing the correctness of the final generated report. The corresponding generated report for a given radiology image is denoted as $\mathbf{y} = (y_1, \cdots, y_n, \cdots, y_N)$, where $y_n \in \mathbb{V}$ is the generated unigram tokens, N is the length of the report, and $\mathbb{V}$ is the vocabulary of all possible v tokens for reports generation. For example, the word sequence "Pleural effusion" is segmented into small pieces of tokens, i.e., {"Pleural", "effus", "ion"}. Generally, the aim of the report generation is to maximise the conditional log-likelihood, i.e.,

$$\theta^* = \arg\max_{\theta} \prod_{n=1}^{N} p_\theta \left(y_n \mid y_1, \ldots, y_{n-1}, \mathbf{I}\right), \tag{1}$$

where θ denotes the model parameters and y_0 represents the start token. After incorporating each disease indicator $\mathbf{c} \in \mathbf{C}$ into the conditional probability $p_\theta \left(y_n \mid y_1, \ldots, y_{n-1}, \mathbf{I}\right)$, we have

$$\log p_\theta \left(y_n \mid y_1, \ldots, y_{n-1}, \mathbf{I}\right) = \int_{\mathbf{C}} \log p_\theta \left(y_n \mid y_1, \ldots, y_{n-1}, \mathbf{c}, \mathbf{I}\right) p_\theta(\mathbf{c} \mid \mathbf{I}) \mathrm{d}\mathbf{c}, \tag{2}$$

where $p_\theta(\mathbf{c} \mid \mathbf{I})$ represents the classifier module.

Recall that our IIHT framework is demonstrated in Fig. 2. The details are described in the subsections below.

3.1 Classifier Module

Image Encoder. The first step in medical report generation is to extract the visual features from the given medical images. In our research, we employ a pre-trained visual feature extractor, such as ResNet [8], to extract the visual features from patients' radiology images that commonly contain multiple view images. For simplicity, given a set of r radiology images $\{\mathbf{I}_i\}_{i=1}^{r}$, the final visual features say $\mathbf{x}$ are obtained by merging the corresponding features of each image using max-pooling across the last convolutional layer. The process is formulated as $\mathbf{x} = f_v \left(\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_r\right)$, where $f_v\left(\cdot\right)$ refers to the visual extractor and $\mathbf{x} \in \mathbb{R}^F$ with F number of features.

Capture Disease Indicator Embedding. The visual features are further transformed into multiple low-dimensional feature vectors, regarded as disease indicator embeddings, which have the capacity to capture interrelationships and correlations among different diseases. The indicator disease embedding is denoted as $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_T) \in \mathbb{R}^{e \times T}$, where e is the embedding dimension and note that T is the number of indicators. Each vector $\mathbf{d}_t \in \mathbb{R}^e, t = 1, \cdots, T$ is the representation of the corresponding disease indicator, which can be acquired through a linear transformation of the visual features, i.e.,

$$\mathbf{d}_t = \mathbf{W}_t^\top \mathbf{x} + \mathbf{b}_t, \tag{3}$$

where $\mathbf{W}_t \in \mathbb{R}^{F \times e}$ and $\mathbf{b}_t \in \mathbb{R}^e$ are learnable parameters of the t-th disease representation.

The intuitive advantage of separating high-dimensional image features into distinct low-dimensional embeddings is that it facilitates the exploration of the relationships among disease indicators. However, when dealing with medical images, relying solely on disease indicator embeddings is insufficient due to the heterogeneous information, including the disease type (e.g. disease name) and the disease status (e.g. positive or negative). Consequently, we undertake further decomposition of the disease indicator embedding, thereby leading to the conception of the subsequent state embedding.

Capture State Embedding. To improve the interpretability of the disease indicator embeddings, a self-attention module is employed to offer valuable insights into the representation of each indicator. Each indicator embedding is further decomposed to obtain the disease state such as positive, negative or uncertain. Let M be the number of states and $\mathbf{S} = (\mathbf{s}_1, \cdots, \mathbf{s}_M) \in \mathbb{R}^{e \times M}$ be the state embedding, which is randomly initialized and learnable. Given a disease indicator embedding vector $\mathbf{d}_t$, the final state-aware of the disease embedding say $\hat{\mathbf{d}}_t \in \mathbb{R}^e$ is obtained by $\hat{\mathbf{d}}_t = \sum_{m=1}^{M} \alpha_{tm} \mathbf{s}_m$, where α_{tm} is the self-attention score of $\mathbf{d}_t$ and $\mathbf{s}_m$ defined as

$$\alpha_{tm} = \frac{\exp(\mathbf{d}_t^\top \cdot \mathbf{s}_m)}{\sum_{m=1}^{M} \exp\left(\mathbf{d}_t^\top \cdot \mathbf{s}_m\right)}. \tag{4}$$

Iteratively, each disease indicator representation $\mathbf{d}_t$ will be matched with its corresponding state embedding $\mathbf{s}_m$ by computing vector similarity, resulting in an improved disease indicator representation $\hat{\mathbf{d}}_t$.

Classification. To enhance the similarity between $\mathbf{d}_t$ and $\mathbf{s}_m$, we treat this as a multi-label problem. The calculated self-attention score α_{tm} is the confidence level of classifying disease t into the state m, which is then used as a predictive value. By abuse of notation, let $\mathbf{c}_t = \{c_{t1}, \cdots, c_{tm}, \cdots, c_{tM}\}$ be the t-th ground-true disease indicator and $\boldsymbol{\alpha}_t = \{\alpha_{t1}, \cdots, \alpha_{tm}, \cdots, \alpha_{tM}\}$ be the prediction, where $c_{tm} \in \{0, 1\}$ and $\alpha_{tm} \in (0, 1)$. The loss of the multi-label classification can be defined as

$$\mathcal{L}_C = -\frac{1}{T} \sum_{t=1}^{T} \sum_{m=1}^{M} c_{tm} \log\left(\alpha_{tm}\right). \tag{5}$$

The maximum value α_{tm} in $\boldsymbol{\alpha}_t$ represents the predicted state for disease t. To enable integration with the indicator expansion module, we adopt an alternative approach; instead of directly utilizing $\hat{\mathbf{d}}_t$, we recalculate the state-aware embedding for the t-th disease indicator, denoted as $\hat{\mathbf{s}}_t \in \mathbb{R}^e$, i.e.,

$$\hat{\mathbf{s}}_t = \sum_{m=1}^{M} \begin{cases} c_{tm} \mathbf{s}_m, & \text{if training phase,} \\ \alpha_{tm} \mathbf{s}_m, & \text{otherwise.} \end{cases} \tag{6}$$

Hence, the state-aware disease indicator embedding $\hat{\mathbf{s}}_t$ directly contains the state information of the disease t.

3.2 Indicator Expansion Module

In the indicator extension module, we employ a "data-text-data" conversion strategy. This strategy involves converting the input indicator embedding from its original format into a textual sequential word representation and then converting it back to the original format. The inherent interpretability of short disease indicator sequences can be further enhanced, resulting in generating more reliable medical reports. For each disease indicator and its state, whether it is the ground-truth label $\mathbf{c}_t$ or the predicted label $\boldsymbol{\alpha}_t$, it can be converted into a sequence of words, denoted as $\hat{\mathbf{c}}_t = \{\hat{c}_{t1}, \cdots, \hat{c}_{tk}, \cdots, \hat{c}_{tK}\}$, where $\hat{c}_{tk} \in \mathbb{W}$ is the corresponding word in the sequence, K is the length of the word sequence, and $\mathbb{W}$ is the vocabulary of all possible words in all indicators. For example, an indicator such as "lung oedema uncertain" can be converted into a word sequence such as {"lung", "oedema", "uncertain"}. To extract the textual information within the short word sequence for each disease t, we use a one-layer bi-directional gated recurrent unit as an encoder say $f_w(\cdot)$ followed by a multi-layer perceptron (MLP) $\boldsymbol{\Phi}$ to generate the indicator information $\mathbf{h}_t \in \mathbb{R}^e$, i.e.,

$$\mathbf{h}_t = \boldsymbol{\Phi}\left(\mathbf{h}_{t0}^w + \mathbf{h}_{tk}^w\right), \quad \mathbf{h}_{tk}^w = f_w\left(\hat{c}_{tk}, \mathbf{h}_{tk-1}^w\right), \tag{7}$$

where $\mathbf{h}_{tk}^w \in \mathbb{R}^e$ is the hidden state in f_w. For each disease indicator, the initial state ($k = 0$) in f_w is the corresponding state-aware disease indicator embedding $\hat{\mathbf{s}}_t$, i.e., $\mathbf{h}_{t0}^w = \hat{\mathbf{s}}_t$.

3.3 Generator Module

The generator say f_g of our IIHT framework is based on the transformer encoder architecture, comprising Z stacked masked multi-head self-attention layers alongside a feed-forward layer positioned at the top of each layer. Each word y_k in the ground-truth report is transferred into the corresponding word embedding $\hat{\mathbf{y}}_k \in \mathbb{R}^e$. For the new word y_n, the hidden state representation $\mathbf{h}'_n \in \mathbb{R}^e$ in the generator f_g is computed based on the previous word embeddings $\{\hat{\mathbf{y}}_k\}_{k=1}^{n-1}$, the calculated indicator information $\{\mathbf{h}_t\}_{t=1}^{T}$ and the visual representation $\mathbf{x}$, i.e.,

$$\mathbf{h}'_n = f_g\left(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_{n-1}, \mathbf{h}_1, \cdots, \mathbf{h}_T, \mathbf{x}\right). \tag{8}$$

For the i-th report, the confidence $\mathbf{p}_n^i \in \mathbb{R}^v$ of the word y_n is calculated by

$$\mathbf{p}_n^i = \operatorname{softmax}\left(\mathbf{W}_p^\top \mathbf{h}'_n\right), \tag{9}$$

where $\mathbf{W}_p \in \mathbb{R}^{e \times v}$ is a learnable parameter and recall that v is the size of $\mathbb{V}$.

The loss function of the generator say $\mathcal{L}_{\mathcal{G}}$ is determined based on the cross-entropy loss, quantifying all the predicted words in all the given l medical reports with their ground truth, i.e.,

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{l} \sum_{i=1}^{l} \sum_{n=1}^{N} \sum_{j=1}^{v} y_{nj}^i \log\left(p_{nj}^i\right), \tag{10}$$

where p_{nj}^i is the j-th component of $\mathbf{p}_n^i$, and y_{nj}^i is j-th component of $\mathbf{y}_n^i \in \mathbb{R}^v$ which is the ground-truth one-hot encoding for word y_n in the i-th report. Therefore, the final loss of our IIHT method is

$$\mathcal{L} = \lambda \mathcal{L}_{\mathcal{G}} + (1 - \lambda) \mathcal{L}_C, \tag{11}$$

where λ is a hyperparameter.

4 Experimental Setup

4.1 Data

The publicly available IU X-Ray dataset [4] is adopted for our evaluation. It contains 7,470 chest X-Ray images associated with 3,955 fully de-identified medical reports. Within our study, each report comprises multi-view chest X-Ray images along with distinct sections dedicated to impressions, findings and indications.

4.2 Implementation

Our analysis primarily focuses on reports with a finding section, as it is deemed a crucial component of the report. To tackle the issue of data imbalance, we utilise a strategy wherein we extract 11 prevalent disease indicators from the dataset, excluding the "normal" indicators based on the findings and indication sections of the reports. Additionally, three states (i.e., uncertain, negative and positive) are assigned to each indicator. In cases where a report lacks information regarding all indicators, we discard the report to ensure data integrity and reliability. The preprocessing of all reports is followed by the random selection of image-report pairs, which are then divided into three sets, i.e., training, validation and test sets. The distribution of these sets is 70%, 10% and 20%, respectively. All the words in the reports are segmented into small pieces by SentencePiece [12]. Standard five-fold cross-validation on the training set is used for model selection.

To extract visual features, we utilise two different models: ResNet-50 [8] pre-trained on ImageNet [5] and a vision transformer (ViT) [7]. Prior to extraction, the images are randomly cropped to a size of 224 × 224, accompanied by data augmentation techniques. Within our model, the disease indicator embedding, indicator expansion module and generator module all have a hidden dimension of 512. During training, we iterate 300 epochs with a batch size of 8. The hyperparameter λ in the loss function is set to 0.5. For optimisation, we employ AdamW [19] with a learning rate of 10^{-6} and a weight decay of 10^{-4}.

4.3 Metrics

The fundamental evaluation concept of the generated reports is to quantify the correlation between the generated and the ground-truth reports. Following most of the image captioning methods, we apply the most popular metrics for evaluating natural language generation such as 1–4 g BLEU [23], Rouge-L [16] and METEOR [6] to evaluate our model.

5 Experimental Results

In this section, we first evaluate and compare our IIHT method with the state-of-the-art medical report generation methods. Then we conduct an ablation study for our method to verify the effectiveness of the indicator expansion module under different image extractors.

Table 1. Comparison between our IIHT method and the state-of-the-art medical report generation methods on the IU X-Ray dataset. Sign † refers to the results from the original papers. A higher value denotes better performance in all columns.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
VTI [20]†	0.493	0.360	0.291	0.154	0.218	0.375
Wang et al. [27]†	0.450	0.301	0.213	0.158	-	0.384
CMR [2]†	0.475	0.309	0.222	0.170	0.191	0.375
R2Gen [3]†	0.470	0.304	0.219	0.165	0.187	0.371
Eddie-Transformer [21]†	0.466	0.307	0.218	0.158	-	0.358
CMAS [10]†	0.464	0.301	0.210	0.154	-	0.362
DeltaNet [29]†	0.485	0.324	0.238	0.184	-	0.379
Ours	**0.513** ± 0.006	**0.375** ± 0.005	**0.297** ± 0.006	**0.245** ± 0.006	**0.264** ± 0.002	**0.492** ± 0.004

5.1 Report Generation

We compare our method with the state-of-the-art medical report generation models, including the variational topic inference (VTI) framework [20], a graph-based method to integrate prior knowledge in generation [27], the cross-modal memory network (CMR) [2], the memory-driven transformer (R2Gen) [3], the co-operative multi-agent system (CMAS) [10], the enriched disease embedding based transformer (Eddie-Transformer) [21], and the conditional generation process for report generation (DeltaNet) [29]. The quantitative results of all the methods on the IU X-Ray dataset are reported in Table 1. It clearly shows that our proposed IIHT method outperforms the state-of-the-art methods by a large margin across all the evaluation metrics, demonstrating the dramatic effectiveness of our method.

The methods under comparison in our study focus on exploring the correlation between medical images and medical reports. Some of these approaches have incorporated supplementary indicators as auxiliary information. However, these indicators primarily comprise frequently occurring phrases across all reports, disregarding the inherent imbalance within medical data. Consequently, the generated reports often treat abnormal patients as normal, since the phrases describing normal areas dominate the dataset. In contrast, our proposed method leverages disease indicators and assigns corresponding states based on the reported content. By adopting a "data-text-data" conversion approach in the indicator expansion module, our method effectively mitigates the issue of misleading the generated medical reports, and thus surpasses the performance of the existing approaches.

5.2 Ablation Study

We now conduct an ablation study for our method to verify the effectiveness of different image extractors. Table 2 presents the results of our experiments, wherein we employed different visual feature extractors with and without the indicator expansion module. Specifically, we exclude the original “data-text-data” conversion strategy; instead, the disease indicator state features are directly used as the input of the MLP layer. This study allows us to analyse the influence of the “data-text-data” strategy within the indicator expansion module on the performance of the proposed IIHT framework.

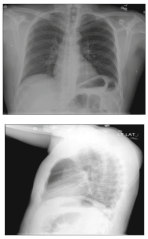
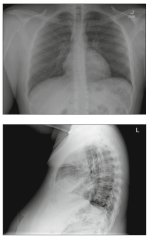
Table 2. The ablation study of our method on the IU X-Ray dataset. “w/o Indicator” refers to the model without the indicator expansion module.

Methods	Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IIHT w/o Indicator	ViT	0.434 ± 0.002	0.294 ± 0.004	0.210 ± 0.004	0.153 ± 0.004	0.216 ± 0.001	0.409 ± 0.005
IIHT (Proposed)		0.463 ± 0.006	0.323 ± 0.005	0.241 ± 0.005	0.186 ± 0.004	0.234 ± 0.003	0.445 ± 0.004
IIHT w/o Indicator	ResNet-50	0.428 ± 0.007	0.271 ± 0.008	0.188 ± 0.003	0.136 ± 0.003	0.185 ± 0.002	0.376 ± 0.004
IIHT (Proposed)		**0.513** ± 0.006	**0.375** ± 0.005	**0.297** ± 0.006	**0.245** ± 0.006	**0.264** ± 0.002	**0.492** ± 0.004

By excluding the incremental disease indicator information, we observe that the image extractor ViT has a better performance than ResNet-50, see the results of the first and third rows in Table 2. This indicates that ViT is capable of effectively capturing semantic feature relationships within images. These findings provide evidence regarding the advantages of ViT in extracting visual information from images. We also observe that utilising indicator information extracted from the indicator expansion module indeed contributes to the generation of precise and comprehensive medical reports, resulting in a noteworthy enhancement in terms of the quality of the generated reports. This improvement is observed when using both ViT and ResNet-50. Interestingly, as indicated in the second and fourth rows in Table 2, when the indicator expansion module is added, the performance improvement of ViT is not as significant as that of ResNet-50. We hypothesise that ViT requires a substantial amount of data to learn effectively from scratch. It is possible that the limited number of iterations during fine-tuning prevents ViT from achieving its full potential in performance enhancement. On the whole, our proposed IIHT method offers significant improvements over the state-of-the-art models. This enhancement can be attributed to the inclusion of the disease indicator expansion module, which plays a crucial role in enhancing the quality of the generated reports.

Finally, in Table 3, we showcase some examples of the reports generated by our method. By incorporating both images and indicators, our method closely mimics the process followed by radiologists when composing medical reports while also addressing the data imbalance challenge. Even in the case where all indicators are normal, a generated report for a healthy patient typically includes

Table 3. Generated samples by our method on the IU X-Ray dataset.

Data	Groud-truth Reports	Generated Reports
	No acute cardiopulmonary findings. No focal consolidation. No visualized pneumothorax. No large pleural effusions . The heart size and cardiomediastinal silhouette are grossly unremarkable.	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. Specifically, no evidence of focal consolidation pneumothorax or pleural effusion. Cardiomediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality.
Indicators: Cardiomediastinal silhouette negative; pneumothorax negative; granuloma negative; consolidation negative; pleural effusion negative; pneumonia negative.		
	Right middle lobe and lower lobe pneumonia. Heart size is within the upper limits of the normal. The pulmonary and mediastinum are within normal limits. there is no pleural effusion or pneumothorax. There is the right basilar air space opacity.	Right lower lobe airspace disease in the right lower lobe atelectasis or pneumonia. Heart size and pulmonary vascularity appear within normal limits. There is no pleural effusion or pneumothorax. There are no acute bony abnormalities.
Indicators: Lung opacity positive; pneumonia positive; pulmonary edema negative; pulmonary negative; pleural effusion negative; and pneumothorax negative.		

a description of various disease indicators, as shown in the first example in Table 3. For patients with abnormal conditions, our method still has a remarkable ability to accurately generate comprehensive reports. Moreover, our method incorporates the capability of facilitating real-time modification of disease indicators, thereby enabling a more accurate and complete process for report generation. This functionality serves to minimise the occurrence of misdiagnosis instances, and thus enhances the overall accuracy and reliability of the generated reports. As a result, we reveal that the generated medical reports with the use of indicator-based features can be more reasonable and disease-focused in comparison to traditional "image-to-text" setups.

6 Conclusion

In this paper, we proposed a novel method called IIHT for medical report generation by integrating disease indicator information into the report generation process. The IIHT framework consists of the classifier module, indicator expansion module and generator module. The "data-text-data" strategy implemented in the indicator expansion module leverages the textual information in the form of concise phrases extracted from the disease indicators and states. The accompanying data conversion step enhances the indicator information, effectively resolving the data imbalance problem prevalent in medical data. Furthermore, this conversion also facilitates the correspondence between the length and correlation of medical data texts with disease indicator information. Our method makes it feasible for radiologists to modify the disease indicators in real-world scenarios and integrate the operations into the indicator expansion module, which ultimately contributes to the standardisation of report fluency and accuracy. Extensive experiments and comparisons with state-of-the-art methods demonstrated the great performance of the proposed method. One potential limitation of our experiments is related to the accessibility and accuracy of the disease indicator information. The presence and precision of such disease indicator information can affect the outcomes of our study. Interesting future work could involve investigating and enhancing our method from a multi-modal perspective by incorporating additional patient information such as age, gender and height for medical report generation.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5904–5914. Association for Computational Linguistics, August 2021. https://doi.org/10.18653/v1/2021.acl-long.459. https://aclanthology.org/2021.acl-long.459
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, November 2020
4. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the 9th Workshop on Statistical Machine Translation, pp. 376–380 (2014)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)
10. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: on exploiting the structure information of chest X-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6570–6580. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1657. https://aclanthology.org/P19-1657

11. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2577–2586. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-1240. https://aclanthology.org/P18-1240
12. Kudo, T., Richardson, J.: SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66–71. Association for Computational Linguistics, Brussels, Belgium, November 2018. https://doi.org/10.18653/v1/D18-2012. https://aclanthology.org/D18-2012
13. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8928–8937 (2019)
14. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
15. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent topic-transition GAN for visual paragraph generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3362–3371 (2017)
16. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
17. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3001–3012. Association for Computational Linguistics, August 2021. https://doi.org/10.18653/v1/2021.acl-long.234. https://aclanthology.org/2021.acl-long.234
18. Liu, F., Ren, X., Liu, Y., Wang, H., Sun, X.: simNet: stepwise image-topic merging network for generating detailed and comprehensive image captions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 137–149. Association for Computational Linguistics, Brussels, Belgium, October–November 2018. https://doi.org/10.18653/v1/D18-1013. https://aclanthology.org/D18-1013
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Najdenkoska, I., Zhen, X., Worring, M., Shao, L.: Variational topic inference for chest X-ray report generation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 625–635. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_59
21. Nguyen, H.T., et al.: Eddie-transformer: enriched disease embedding transformer for X-ray report generation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
22. European Society of Radiology (ESR) communications@myesr.org: Medical imaging in personalised medicine: a white paper of the research committee of the European society of radiology (ESR). Insights Imag. **6**, 141–155 (2015)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

24. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2506 (2016)
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
26. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
27. Wang, S., Tang, L., Lin, M., Shih, G., Ding, Y., Peng, Y.: Prior knowledge enhances radiology report generation. In: AMIA Annual Symposium Proceedings, vol. 2022, p. 486. American Medical Informatics Association (2022)
28. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: TieNet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9049–9058 (2018)
29. Wu, X., et al.: DeltaNet: conditional medical report generation for COVID-19 diagnosis. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 2952–2961. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, October 2022. https://aclanthology.org/2022.coling-1.261
30. Yin, C., et al.: Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 728–737. IEEE (2019)
31. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016)
32. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8739–8748 (2018)