Towards Automated and Accurate Radiology Report Generation

by

Tran Nhat Hoang Nguyen

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Signal and Image Processing

Department of Electrical and Computer Engineering

University of Alberta

Abstract

Radiology reports are the primary medium through which physicians communicate findings and diagnoses from patients' medical scans. Examples include radiology reports for chest radiographs, CT scans of the brain, medical reports of retinal images, and more. However, the process of writing medical reports is tedious, error-prone, and time-consuming, even for experienced radiologists. Moreover, a Covid-19 or similar pandemic could exacerbate the existing problems to all health care systems worldwide. Therefore, this thesis explores the ability to automate diagnosing diseases and accurately generate radiology reports to alleviate the burdens of medical doctors.

This thesis describes a new fully end-to-end differentiable paradigm that consists of three major complementary modules: Classifier, Generator, and Interpreter. Particularly, taking the chest radiographs and related information as inputs, the *classifier* module produces state-aware disease embeddings by polarizing visual disease features into different directions, referred to as disease states (e.g., positive, negative, uncertain, or unmentioned). With the awareness of the disease states, a semantic version of the disease representation is formed, referred to as EnricheD DIsease Embeddings (EDDIE), and passed to a transformer-based *generator* to produce meaningful medical reports. The generated reports are fed to the *interpreter* to ensure consistency with respect to the *disease classification checklist*. This three-step approach ensures that the visual information is always semantic enough to generate medical reports. Meanwhile, the generated reports must exactly describe the detected diseases,

avoiding overfitting to any dominant class (e.g., due to imbalanced datasets) or language metric (i.e., by cheating the generation process).

The proposed model is evaluated on different datasets with commonly-used metrics concerning language fluency, clinical accuracy, and human evaluation. Empirical evaluations demonstrate that the proposed model can make more accurate diagnoses and generate more fluent and precise reports than existing baselines. Moreover, noticeable performance gains are consistently observed when additional contextual information is available, such as the patients' clinical background documents and extra scans from different views.

Preface

Some of the research conducted for this thesis forms part of an international research collaboration, led by Prof. Li Cheng at the University of Alberta, with Dr. Dong Nie being the lead collaborator at the University of North Carolina at Chapel Hill, and Prof. Yingying Zhu at the University of Texas at Arlington. The methods in Chapter 3 and Chapter 4 were designed by myself, with the assistance of Dr. Dong Nie. The data analysis in Chapter 3 and Chapter 4 and concluding analysis in Chapter 5 are my original work, as well as the literature review in Chapter 2. The human evaluation in Chapter 3 and Chapter 4 are done by Dr. Yujie Liu at Guangzhou University of Chinese Medicine.

Chapter 3 of this thesis is under review at *The IEEE International Symposium on Biomedical Imaging (ISBI 2022)*.

Chapter 4 of this thesis has been published as Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. "Automated Generation of Accurate & Fluent Medical X-ray Reports." *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3552-3569. 2021.

This research study was conducted using human subject data made available in open access [9], [10], [29], [67]. Ethical approval was not required, as confirmed by the license attached with the open-access data.

To my mom, dad, brother, and beloved wife
For supporting me in everything.

If the hand be held between the discharge-tube and the screen, the darker shadow of the bones is seen within the slightly dark shadow-image of the hand itself... For brevity's sake I shall use the expression "rays"; and to distinguish them from others of this name I shall call them "X-rays".

– Wilhelm Röntgen, German physicist, 1895.

Acknowledgements

I want to send my sincere gratitude to Prof. Li Cheng for his continuous support and discussions. Prof. Li Cheng gave me a fantastic opportunity to participate in a world-class research journey. I would also like to thank Dr. Dong Nie for his excellent guidance and collaboration to help me develop many ideas. His research and life experience helped me during my most stressful and challenging time. In addition, I would like to thank all members of Vision and Learning Laboratory for their well-prepared research presentations every week. Special thanks to Shihao Zou, Ji Yang, Taivanbat Badamdorj, Chuan Guo, Mahdiar Nekoui, and Javad Khaghani.

I also want to thank all lecturers in the Department of Electrical and Computer Engineering and Computing Science. They have helped me by teaching unforgettable lessons to solidify my background in Machine Learning, Computer Vision, and related areas.

Finally, I would like to thank my family: my mom, dad, brother, and beloved wife. Though separated by distance, they never make me feel alone. Thank you so much for your great sacrifices. For my wife, thank you for being with me for almost one year to take care of me, especially during the pandemic breakout. And, I am very proud of you that you are now a Ph.D. student. It's been amazing to see your growth and maturity in your research career.

Contents

List of Tables

List of Figures

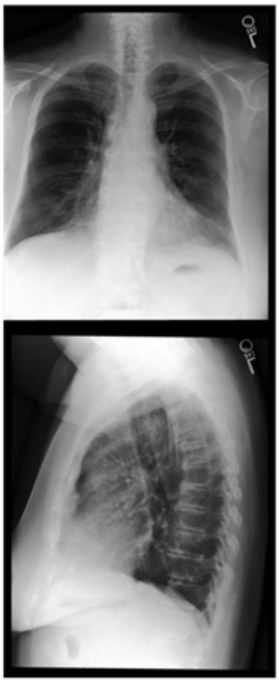
Chapter 1

Introduction

Radiology reports play a pivotal role in the clinical environment. It covers some background information of patients and disease findings written by radiologists. It evaluates conditions of patients based on visual observation and serves as a medium to communicate between doctors and patients for medical treatment discussions [70].

However, to write an excellent medical report [20], it usually requires a considerable amount of time to review patients' clinical history, understand the doctor's indications [48], and investigate medical scans to identify abnormalities. After that, radiologists write down their findings into the reports describing diseases with condensed sentences [70] (see Figure 1.1). Notably, although some diseases exist in medical scans, radiologists have to prioritize crucial diseases that address the current concerns of medical doctors [45]. It can be thought of as a question answering problem where only the needed information is displayed in the final radiology reports. Therefore, the whole process becomes tedious and time-consuming for experienced radiologists and error-prone for junior ones.

Meanwhile, the Covid-19 pandemic has been putting enormous pressure on hospital systems worldwide. Physicians are working excessively to treat Covid-19 patients leading to labor shortage [12] to treat other diseases. Although the Covid-19 rapid antigen test kits can detect new cases quickly and reliably, they cannot determine the severity or be able to guide treatment and assess treatment response [6]. So far, screening infected patients with radiology



EXAM: Chest frontal and lateral views.

CLINICAL INFORMATION: Mechanical fall from standing, on Coumadin. Neck pain.

COMPARISON: None.

FINDINGS: Frontal and lateral views of the chest were obtained. There is mild basilar atelectasis without evidence of focal consolidation. No pleural effusion or pneumothorax is seen. There is minimal biapical pleural thickening. Cardiac silhouette is top normal with likely adjacent epicardial fat pad. The aorta is calcified and tortuous. Some degenerative changes are seen along the spine.

IMPRESSION: No acute cardiopulmonary process.

Figure 1.1: An example of a radiology report study.

examinations using chest radiography has been the most effective approach, emphasizing its crucial role in modern clinical systems. However, it also exposes another bottleneck issue where all available resources are prioritized to fight this disease leading to a longer waiting time for other treatments [62].

These challenges indicate that an effective clinical system must resolve the labor shortage problem and reduce the time needed to diagnose or assess diseases using radiography. It has led to a surging need for automated generation of medical reports to assist radiologists and physicians in making rapid and meaningful diagnoses [27], [28], [33], [34]. Its potential efficiency and benefits could be enormous, especially during critical situations.

This thesis focuses on developing an automated system that can quickly and accurately diagnose diseases and generate meaningful radiology reports to alleviate the burdens of physicians and radiologists.

1.1 Challenges

Although having an automated medical report generating system is essential, developing a robust system is highly challenging with a growing number of related research works. Notably, many existing efforts primarily focus on image-based captioning problems by summarizing visual information in images or videos with a sentence or a topic-related paragraph [15], [17], [24], [51], [55], [61], [65], [73]. Based on these works, many other works have been adjusted to meet the specific needs for medical report generation [16], [23], [27], [28], [33], [34], [40], [47], [58], [59], [68], [74]–[76], [78], [81]. Compared to regular image captioning problems where there is no need for expert knowledge, medical report generation is much more difficult [28] because medical images usually lack rich contextual information and are more difficult to understand [56]. Moreover, each radiology study is domain-specific and precise, where each report typically consists of many long sentences describing different disease findings. Thus, the main challenges associated with radiology report generation lie in:

- Accurately recognizing diseases from medical scans (factual correctness).
- Describing the disease findings with correct medical terminology (language fluency).
- Handling potential inconsistencies between the detected diseases and generated reports due to long-range dependency issues (clinical coherence).

1.2 Contributions

The aforementioned observations motivate us to propose a categorize-generate-interpret framework that places specific emphasis on clinical accuracy while maintaining adequate language fluency of the generated reports. The main contributions of this work are:

- In Chapter 3, we present an EnricheD DIsease Embedding based Transformer (Eddie-Transformer) model, which jointly performs disease de-

tection and medical report generation. This is done by decoupling the latent visual features into semantic disease embeddings and disease states via our state-aware mechanism. Then, our model entangles the learned diseases and their states, enabling explicit and precise disease representations.

- Based on Chapter 3, in Chapter 4, we propose a fully differentiable and end-to-end paradigm that contains three complementary modules: taking the chest X-ray images and clinical history documents of patients as inputs, our *classification* module produces an internal checklist of disease-related topics, referred to as enriched disease embedding; the embedding representation is then passed to our transformer-based *generator*, to produce medical reports; meanwhile, our generator also creates a weighted embedding representation, which is fed to our *interpreter* to ensure consistency with respect to disease-related topics.

1.3 Thesis Overview

In Chapter 2, we review existing efforts, their limitations, and other related works. In Chapter 3, we present a new method to detect diseases and enhance disease feature representations to facilitate the radiology report generation task. In Chapter 4, we consider a broader range of inputs and propose a robust model to diagnose, generate, and fine-tune medical reports. Finally, in Chapter 5, we summarize and conclude this thesis and discuss possible future research directions.

Chapter 2

Related Work

This chapter discusses the works of medical report generations by first reviewing existing efforts in the image-captioning problem and their limitations. After that, this chapter focuses on some radiology report methods and their advantages and disadvantages. Moreover, this chapter introduces some related techniques involving the natural language processing area. Finally, this chapter summarizes and outlines the limitations and research gaps and how report generation approaches can be improved.

2.1 The Image Captioning Task

The image-based captioning task aims at generating realistic sentences or topic-related paragraphs to summarize visual contents from images (see Figure 2.1) or videos. It is similar to the medical report generation task in that both receive images and generate descriptive sentences explaining the images. Hence, it is reasonable to assume that solving the image captioning problem can help the medical report generation task later. This section starts with how an image captioning problem can be solved, discusses the limitations of the current approaches, and finally how a new model can replace the existing methods in resolving the limitations.

2.1.1 The Recurrent Neural Network Models

The work of [65], [73] are among the earliest and most influential approaches that can, for the first time, address the image captioning problem in an end-

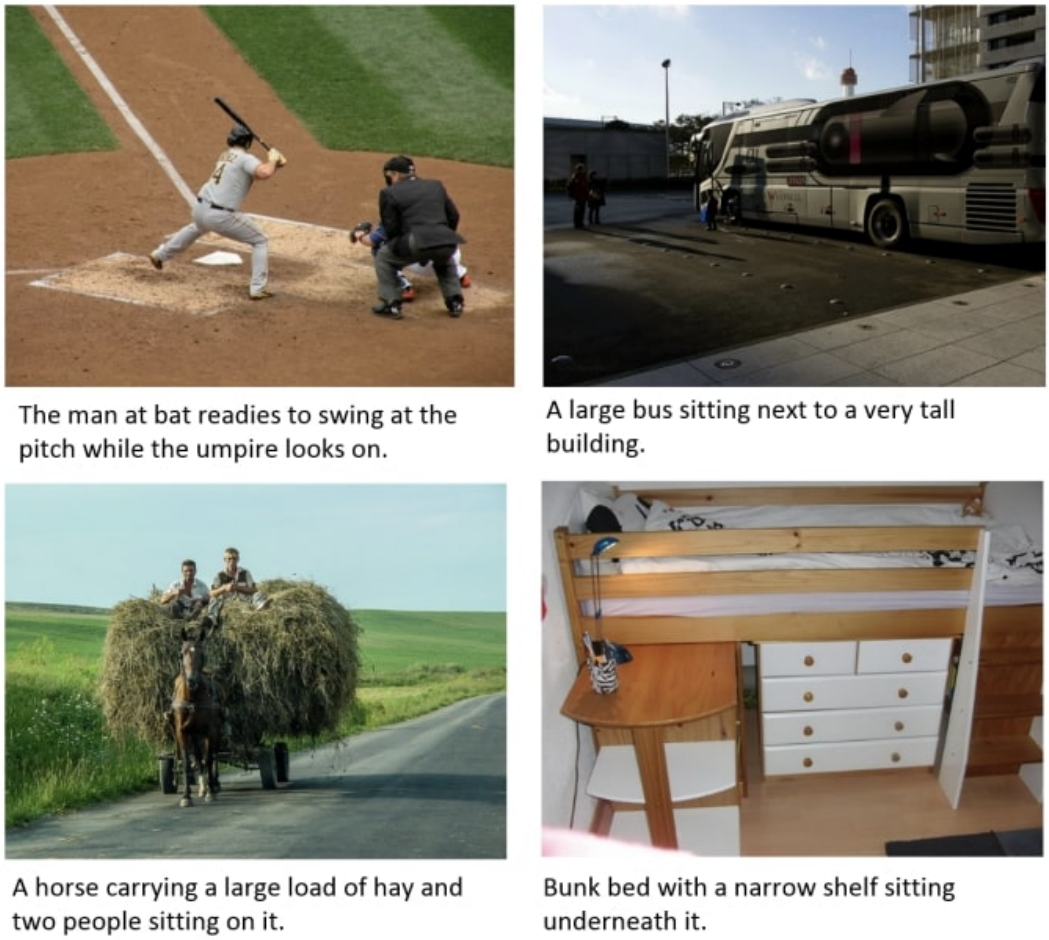


Figure 2.1: Example images with captions obtained from the MSCOCO caption dataset [8].

to-end manner. Particularly, [65] used a CNN image encoder to encode images into latent visual features. These latent visual features are then used as initial inputs for the RNN/LSTM models to generate words (see Figure 2.2). Based on the work of [65], [73] tries to learn the relationship between a generated word and the spatial pixel features from an image via an attention mechanism. Thus, it enables the ability to visualize how deep neural networks describe images and increase the quality of generated captions.

Their experiments evaluated on the MSCOCO dataset [37], which consists of more than 100,000 images, indicate that the proposed CNN-RNN architecture [65] (scoring 27.7 in BLEU-4 metric) can closely match with real human annotators (scoring 21.7 in BLEU-4 metric). The results are even better with

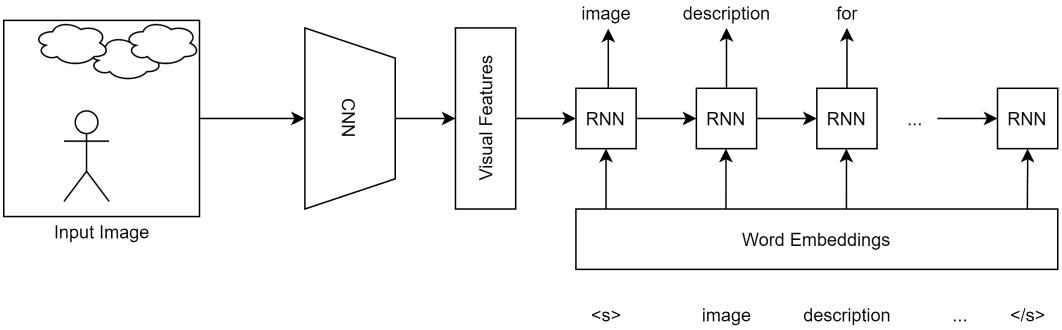


Figure 2.2: An illustration of how CNN-RNNs image captioning model works.

attention mechanism [73] where it outperforms the work of [65] consistently across all language metrics with 2-5% improvement. As a result, they set a strong foundation for many other related works such as visual question answering [17], improving the existing image captioning LSTM models [24], [55], unsupervised image captioning [15], video captioning [51], or medical report generation [28].

2.1.2 The Existing Problems

Despite the enormous successes of [65], [73], most works in this area share the same recurrent neural networks with a widely known long-range dependency issue [14]. For instance, [31] shows that the RNN-based architecture has a bottle-neck issue by modeling all generated words and image features with a single high dimensional embedding. This embedding is then used as input to generate the next word. When the generated sentence is too long, the gradient flows through this network is either too small (vanishing gradient) to significantly impact the learning process or too large (exploding gradient) that can ruin the entire learning process [21]. Although the LSTM model was introduced to bypass this problem by forgetting unrelated words and retaining only the most relevant words, in practice, it is still struggling to generate long paragraphs or documents.

To mitigate this problem, in 1996, [14] proposed a hierarchical recurrent network that transforms a time-series sequence into multiple summarized parts such as a stream of words forming a sentence or multiple sentences forming a

paragraph. Almost ten years later, [31] came up with a similar solution, designed specifically for the image captioning task. It is done by generating sentence embeddings from images with a sentence-level RNN module, then generating words for each sentence using a word-level RNN. Experiments of [31] show that their model can match a real human in describing images with complex paragraphs, achieving 41.9 in BLEU-1 score compared to the human ability with a score of 42.88. Unarguably, the recent progresses in medical report generation [27], [28], [33], [68], [74], [75], [78], [81] have been particularly influenced by the work of [31].

It is clear that hierarchical RNNs [14], [31] can solve the long-range dependency limitation, which helps generate paragraph-level contents. However, as more and more data becomes available, training the hierarchical RNN model can be extremely slow [63] where the time complexity to go through the entire sequence is $O(n)$. It is because the RNN itself is a sequential model. It has to wait for the previous output to be used as input to generate the subsequent output, thus, limiting the parallelism. Apparently, if parallelism is the top requirement, RNN architectures should be avoided in the first place due to their sequential nature.

2.1.3 The Transformer Model

The existing problems discussed in the previous part suggest that a new model must be proposed to solve the long-range dependency and parallelism simultaneously. As it seems impossible because all prior works [14], [31], [44], [65], [73] are based entirely on the RNNs, [73] left a crucial piece in solving this problem – the attention mechanism [44].

The Transformer model [63] is firstly introduced in the context of machine translation (sequence-to-sequence) to expedite training and improve long-range dependency modeling (see Table 2.1). It processes sequential data in parallel with an attention mechanism, consisting of a multi-head attention module and a feed-forward layer. With the proposed Transformer model, recent transformer-based models have shown considerable advancement in many difficult tasks, such as a graph attention network [64], image generation [7], story

Methods	Sequential operations	Maximum path length	Complexity per layer
Transformer [63]	O(1)	O(1)	$O(N^2 * D)$
RNNs	O(N)	O(N)	$O(N * D^2)$

Table 2.1: The complexity summary of the Transformer model [63], where N is the sequence length and D is the embedding dimension

generation [52], question answering, or language inference [11].

Although the Transformer model has been proved to efficiently resolve most limitations of the RNN/LSTM architectures with faster training time [79], very few works [61] have successfully utilized it for the image captioning task, leaving a potential research gap for future works. It is even more interesting to see how the Transformer can be applied to solve real-world problems such as medical report generation, which requires absolute correctness. Therefore, this thesis explores the feasibility of developing such models for medical report generation.

2.2 The Medical Report Generation Task

Despite sharing some similarities with the image captioning task, the medical report generation task is usually more difficult [28] since medical images usually lack rich contextual information and are more difficult to understand. Moreover, each medical report is very domain-specific, precise, and typically consists of many long sentences describing exactly the severity and details of diseases [70]. For these reasons, some research works [28], [33] find that regular image captioning approaches cannot be applied directly to this problem. There are many studies [27], [28], [33], [68], [74], [75], [78], [81] have been proposed to adapt to this specific type of problem, which can be divided into two groups:

- Image-based report generation: The nature of these methods is very similar to the image captioning task of [31]. It receives medical images and generates medical reports hierarchically.

- Template retrieval and paraphrasing: These methods often leverage human knowledge to preprocess information, group medical sentences or

reports into categories. Then, they treat the problem as a template retrieval task and paraphrase the retrieved templates where necessary.

This section discusses common assumptions of these approaches, limitations of existing efforts and introduces an automated evaluation tool to evaluate the generated reports.

2.2.1 The common assumptions

Almost all early medical report generation approaches [27], [28], [38], [68], [75] are entirely based on the assumption that writing good medical reports only needs single-view or frontal-view images, which are captured front-to-back or back-to-front. This assumption is aligned with the common assumption made by physicians in some studies [4], [49]. It has the benefit of reducing the amount of radiation on patients. However, other groups of physicians [2], [26], [30] find that using other view positions such as lateral view has a complementary effect on detecting certain diseases that are often missed out with frontal view images. This thesis agrees on the views of both sides in a way that deep learning models must be flexible enough to handle not only single-view images but also multi-view images when available.

In reality, having images alone may not be good enough to write good medical reports. Human radiologists are much more flexible and often require more contextual information to make precise diagnoses. For example, [5] suggests that the clinical information of patients can help improve the accuracy of diagnoses. In particular, clinical history contains personal information, such as age, gender, and relevant medical information. It may also include previous clinical studies of patients, known diseases, and symptoms the patients may be feeling. Although it is a crucial piece of information for radiologists to focus the report on unique conditions, to our knowledge, none of the prior works has utilized this piece of information. Notably, this information is abundantly provided in most X-ray datasets such as [10], [29].

This thesis (see Chapter 4) shows that deep learning models are flexible enough to handle both cases efficiently with consistent improvements. Thus,

it proves the necessity of having more contextual data in diagnosing diseases.

2.2.2 Image-based Report Generation

The work of [65], [73] are among the early baselines that researchers [28], [33], [38] apply in the medical report generation task. The visual features are extracted by convolution neural networks (CNNs); then, they are subsequently fed into recurrent neural networks (RNNs) to generate textual descriptions. By using these models, researchers imply that visual disease patterns can be learned implicitly via the image captioning process. However, these standard approaches have two major drawbacks.

One is the long-range dependency issue, where each medical report is a paragraph, as discussed earlier. Thus, [27], [28], [33], [68], [74], [75], [78], [81] adopt the hierarchical RNN architecture [31] in their models to deal with this issue by generating one sentence at a time. The only issue with this strategy is the lack of parallelism, affecting the training time [79].

The other one is the inaccurate textual descriptions [33], [38]. One of the reasons behind this is the imbalanced X-ray datasets [10], [67], where the normal study cases dominate the abnormal cases by a ratio of at least 10:1. Therefore, the deep learning models can cheat the learning process by only generating healthy sentences with a high language metric scores. To remedy this issue, some researchers [28], [59] introduce a secondary classification task to detect diseases in conjunction with adjusting the learning weight towards the positive cases [53], enhancing the quality of the visual features. This approach can be thought of as an explicit disease diagnosis process with clear predictions, forcing the report generator to generate what the visual extractor sees. The methods of [27], [33], on the other hand, consider a reinforcement learning process to promote generating reports with correct contents. The downside of reinforcement learning is the notorious convergence difficulty and hard to train.

2.2.3 Template Retrieval & Paraphrasing

Instead of generating medical reports from images, another exciting direction is trying to query or retrieve medical templates and paraphrasing them into a complete medical report [33], [34]. It is achieved by leveraging human knowledge and manual effort to group sentences or medical reports into different categories, such as grouping by similar meaning or diseases [33], [34].

Compared to the approaches in the previous part, this solution can reduce the need for RNN/LSTM models to generate long sequences, thus avoiding the long-range dependency issue. Moreover, using medical templates has transformed this approach into sequence-to-sequence modeling where recurrent models can do their job efficiently by paraphrasing a template sequence into a complete sentence based on visual information.

Although [34] is currently one of the state-of-the-art approaches, it has several noticeable limitations. Firstly, it requires human efforts [33], [34] to establish a database of templates manually. It leads to the second limitation; it is not scalable to other related problems. Each medical dataset would require human effort to scan through the entire dataset and extract templates. For small-scale datasets such as Open-I [10] with less than 4000 studies, it is feasible. Still, with enormous datasets such as MIMIC-CXR [29] with more than 100,000 studies, it is very challenging. Moreover, if a particular template for a specific disease is missing, the model may not generate a desirable medical report. This is not to mention that each institution or country may have its convention or style regarding how to write medical reports [70].

2.2.4 CheXpert Labeler

The CheXpert labeler [25] is a rule-based system that extracts and classifies medical reports into 14 common diseases. Each disease label is either positive, negative, uncertain, or unmentioned. This is a crucial part of building large-scale chest X-ray datasets, such as [25], [29], where an alternative manual labeling process may take years of effort. It can also be used to evaluate the clinical accuracy of a generated medical report [38]. Another important

use of the CheXpert labeler is to facilitate the generation of medical reports. Since the rule-based CheXpert labeler is not differentiable, it is regarded as a score function estimator for reinforcement learning models [38] to fine-tune the generated texts. However, the reinforcement learning methods are often computationally expensive and practically difficult to convergence. As an alternative, Lovelace et al. [40] propose an attention LSTM model and fine-tune the generated report via a differentiable Gumbel random sampling trick, with promising results.

2.3 Multi-label Classification

Unlike multi-class classification problems where each input sample can have one and only one output class, multi-label classification (MLC) assigns each sample a set of target labels. For example, most real-world images contain multiple labels, which could correspond to different objects, scenes, actions, and attributes in an image. The most straightforward MLC approach is transforming the problem into independent binary classification tasks where the final layer consists of independent logistics activation. Then, cross-entropy loss or ranking loss [35], [69] can be applied to train neural network models. Additionally, to capture the relationships between different labels, [66] proposed to learn semantic relevance between images and labels by joining both image embedding and label embedding into a joint embedding space through recurrent neural networks.

Inspired by the work of [66], our model learns to cluster input images into one of the predefined states where each state is an embedded vector. However, unlike existing clustering techniques that form groups of data points, our model clusters visual features into different directions.

2.4 Natural Language Processing Techniques

Natural Language Processing (NLP) is a branch of computer science and machine learning that process, understand, and respond to textual information. In recent years, NLP has been one of the most fast-growing fields with many

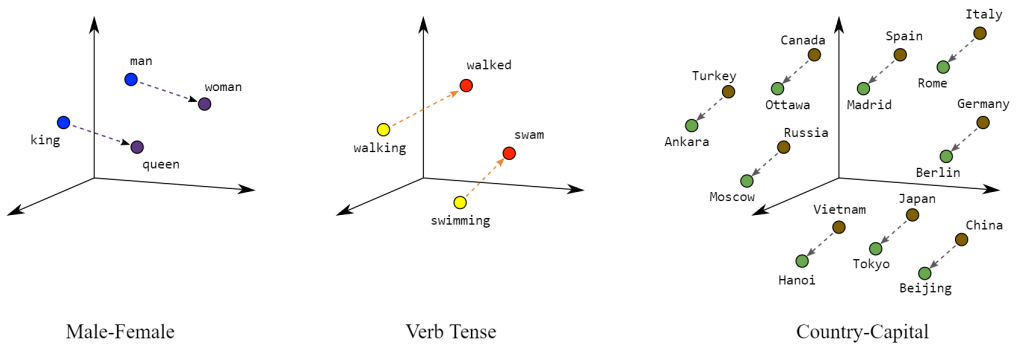


Figure 2.3: An example of word embedding. Image Source: (Embeddings: Translating to a Lower-Dimensional Space) by Google.

exciting applications such as language translation (sequence-to-sequence modeling), image captioning (image-to-sequence modeling), text prediction (time-series forecasting). Since this thesis involves the NLP research areas, this section provides some background knowledge and recent NLP techniques.

2.4.1 Word Embedding & Positional Encoding

The first and most crucial step in processing textual information is transforming a word into a continuous space vector representation [46] which is known as word embedding (see Figure 2.3). It is used to input a human-readable word such as English or French into a machine learning model that it can understand. Word embedding can be used to visualize words in a high-dimensional embedding space where similar words such as "beautiful", "stunning", "awesome" tend to cluster together to form certain relationships.

Another interesting embedding that recently gained popularity is positional encoding. Positional encoding is used to insert time-series information in the Transformer model [63]. This is because the Transformer does not have the sequential characteristic that recurrent models such as LSTM have. For example, the sentence "cat eats fish" without ordering has an equivalent meaning as "fish eats cat". Fortunately, there are many ways to encode positional information such as sinusoid [63], learnable positional encoding [11], or relative positional encoding [71].

2.4.2 Tokenizers

In the previous part, we discussed word embedding, which turns a word into a vector representation. In this part, we discuss tokenization which is an important process that turns a sentence into words [18] or sub-words [57] before transforming words into word embedding. Particularly, a sentence often contains many words, and each word can be associated with punctuations or special characters. For example, a sentence "Washington isn't in the U.K." can be tokenized into "Washington", "isn't", "in", "the", and "U.K.". However, the word "isn't" can be further tokenized into "is" and "n't", whereas "U.K" cannot be broken down further.

Word Tokenizer with Spacy

Spacy is a library that is introduced to facilitate the tokenization task. It first split raw text with whitespace characters. On each substring, it performs two checks:

- Predefined exception rules: there are special rules in English that can be tokenized further, such as the word "don't" should be split into two tokens, "do" and "n't", while "U.S." should remain one token.
- Punctuation check: punctuation such as commas, periods, hyphens, or quotes can be used to split a word into prefix, postfix, and infix, where each of them will be checked. For example, "Eddie-transformer" can be split further into "Eddie","-", and "transformer".

In Chapter 3, we use Spacy for word tokenization. Moreover, we also use Spacy to split sentences into noun phrases to get more additional labels to increase the difficulty of the image classification task.

Other Tokenizers

Although word tokenization is frequently used, it may run into the "out-of-vocabulary" situation where a new word is not in the predefined vocabulary set. Therefore, some other tokenization techniques are introduced, which break

a word into subwords such as Byte Pair Encoding (BPE) [57], Unigram, or WordPiece. These techniques surely avoid the "out-of-vocabulary"; however, the input sentence length may increase significantly. It would be extremely challenging to tackle a paragraph-level problem such as medical report generation, especially for RNN-based architectures, as discussed previously.

In Chapter 4, we use a hybrid approach that keeps a high-frequency word list while reserving a small portion of the vocabulary size for sub-word level tokenizers. With this technique, we balance the sequence length and avoid the out-of-vocabulary situation. To do this, we use SentencePiece library [32].

2.4.3 Multi-head Attention

The multi-head attention (MHA) mechanism was introduced in the Transformer model [63]. Since then, it has been widely applied to many tasks such as graph attention networks (GAT) [64] or image feature extraction via split attention (ResNeSt) [80]. The multi-head attention is based on the information theory with three technical terms: query, key, and value. The query is a learnable embedding that queries information from an input text sequence (database) such that it matches with specific keys or words that share similar meanings. The matching score is computed by a vector similarity dot product between the query embedding and any available key embeddings. Unlike in database query, where only the matched value is returned, the returned value in MHA is a weighted average of all value embeddings. For example, in the sentence "the patient has cough and shortness of breath", the words "cough" and "shortness of breath" are two keywords that may match the "fever" query. Therefore, the word embedding of "cough" and "shortness of breath" are returned by averaging all word embedding in the sentence where "cough" and "shortness of breath" receive the highest weight. This ensures that the outputs only retain the most relevant words. This thesis uses the attention mechanism in many tasks, such as summarizing medical reports based on different diseases where each disease is a query.

2.4.4 Language Evaluation Metrics

BLEU Scores

BLEU (BiLingual Evaluation Understudy) [50] metric was developed with a purpose of automatically evaluating machine-translated text. The BLEU score is a number between zero (no overlap/low quality) and one (perfect overlap/high quality) that measures the similarity of the machine-translated text (i.e., candidate) to a set of high quality reference translations (i.e., references). The BLEU scores can be computed using the following formula:

$$BLEU = min(1, \exp(1 - \frac{r}{c})) \times \exp(\Sigma_{i=1}^{N} w_i log(p_i)) \tag{2.1}$$

$$p_i = \frac{\Sigma_{C \in can} \Sigma_{gram_n \in C} Count_{clip}(gram_n)}{\Sigma_{C' \in can} \Sigma_{gram'_n \in C'} Count(gram'_n)} \tag{2.2}$$

where r and c are the length of the reference and candidate sequences, respectively; w_i is the weight.

ROUGE Scores

Similar to BLEU scores, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [36] metric is another metric used to evaluate a generated text. It is developed with a purpose of evaluating automatic summarization and machine translation. ROUGE score can be computed as follow:

$$ROUGE = \frac{\Sigma_{S \in ref} \Sigma_{gram_n \in S} Count_{match}(gram_n)}{\Sigma_{S \in ref} \Sigma_{gram_n \in S} Count(gram_n)} \tag{2.3}$$

2.5 Conclusion

In summary, although the regular image captioning problem shares some similarities with the medical report generation task, it is much more challenging to tackle the report generation without any modification. Despite having different solutions ranging from detecting diseases, fine-tuning generated reports, and paraphrasing available templates, each approach has its disadvantages, including the imbalanced datasets, parallelism limitation, convergence difficulties of reinforcement learning, and scalability across different datasets. Moreover, the assumptions made by prior works excluding the necessity of having more

contextual data also restrict them from practical usage and do not reflect the actual clinical environment.

From the limitations and research gaps mentioned above, this thesis aims at resolving the long-range dependency issue and parallelism with the Transformer model. This thesis also designs modules such as multi-view image encoders and text encoders to adapt to different scenarios where various clinical information is available such as single-view versus multi-view images, clinical history, or indications. It also improves the quality of the generated reports in terms of clinical correctness by enhancing the visual input features and fine-tuning the generated reports with the proposed differentiable evaluation network called the interpreter.

Chapter 3

Eddie-Transformer: EnricheD DIsease Embedding Transformer for X-ray Report Generation

3.1 Introduction

Physicians primarily communicate findings and diagnoses from patients' medical scans through radiology reports. However, the process is laborious and error-prone, where typing out a medical report typically takes five to ten minutes [28]. In COVID-19 or similar pandemics, a rapidly soaring number of patients could bring enormous pressure to the healthcare system, a devastating setting that calls for automation to lessen healthcare workers' burden. More specifically, we look at the problem of automated generation of medical reports that facilitate rapid and meaningful diagnoses and save time during a critical situation.

Although many medical report generation approaches [16], [27], [28], [33], [34], [47], [58], [59], [68], [74]–[76], [78], [81] have been proposed, many existing works are based on the CNN-RNN paradigm that suffers from poor long-range dependency modeling capability. It also recurrently processes the sequential information limiting the training process and thus leading to sub-optimal results for the sequential description generation. The Transformer [63], which is developed to beat the drawbacks of RNN architectures, has achieved great

success in natural language processing tasks because of its parallelization and attention mechanisms. However, very few works have used Transformer-based models for the medical report generation task. The nice characteristics of the Transformer motivate us to explore the feasibility of developing a new paradigm for image description generation to overcome the restrictions of the CNN-RNN paradigms.

This chapter introduces a simple end-to-end medical report generation framework, EnricheD Disease Embedding Transformer (Eddie-Transformer). The model consists of three modules: a visual feature extractor CNN, an enriched disease embedding block (Eddie), and a report generation transformer. The visual feature extractor module extracts global information of the images, which outputs the visual features. The enriched disease embedding module decouples the visual features to disease query embedding and disease-state (e.g., positive, negative, uncertain), then learns a state-aware disease embedding via a self-attention mechanism. Finally, the report generation module generates reports via the Transformer model based on the enriched disease embedding. The main findings in this chapter are:

- We propose an Enriched Disease Embedding module that has the ability to polarize visual disease features; thus, it encodes informative disease features via the self-attention mechanism, improving the quality of medical reports.
- As a by-product of this polarization process, Eddie has the ability to perform classification tasks.
- The Transformer model can be used to generate medical reports, thus reducing memory consumption compared to LSTM models and faster training time (see Table.3.4).

3.2 Our Approach

Our model first extracts the latent visual features from the last layer of a CNN image encoder. The latent features are then transformed into differ-

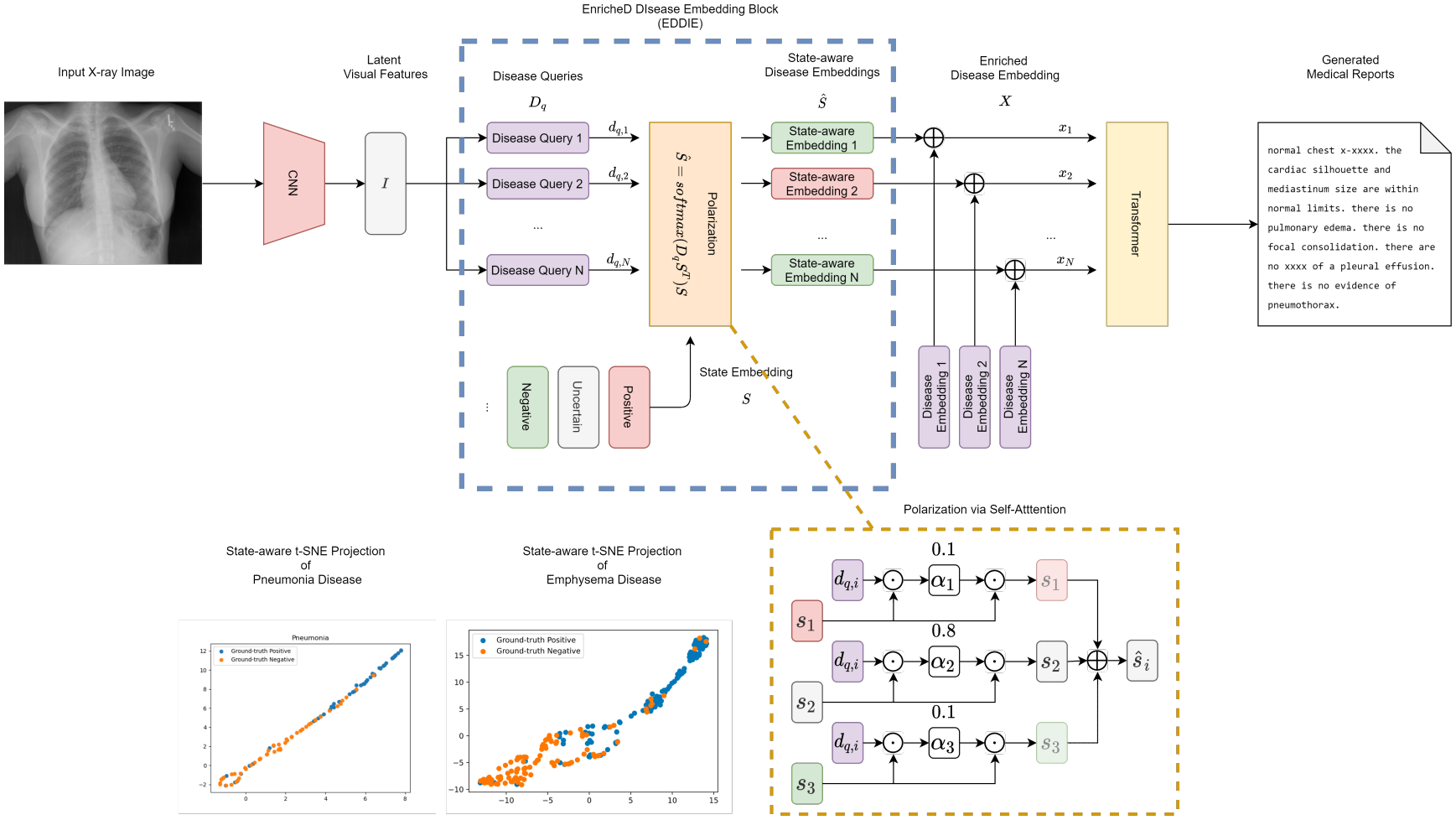


Figure 3.1: The proposed EnricheD DIsease Embedding based Transformer (Eddie-Transformer) model. It consists of three main components: CNN-backbone, enriched disease embedding block, and report generation Transformer.

ent disease/topic queries. This is followed by the EDDIE block that returns the corresponding state-aware disease embedding for each query. Finally, the language model generates medical reports based on the enriched disease embeddings as visualized in Fig. 3.1.

3.2.1 Enriched Disease Embedding

Naive Disease Query Embedding

Denote N the total number of disease representations, and E the embedding dimension. Fig. 3.1 presents the global visual features I extracted from the CNN backbone. Disease query representations are subsequently obtained by transforming the image feature $I \in \mathbb{R}^F$ into multiple low dimensional feature vectors. They are regarded here as the *disease queries*: $\{d_{q,i}\}_{i=1}^{N} \in \mathbb{R}^E$, by a linear form of I ($\phi_i(I)$),

$$d_{q,i} = \phi_i(I) = W_i^T I + b_i, \tag{3.1}$$

Here $W_i \in \mathbb{R}^{F \times E}$ and $b_i \in \mathbb{R}^E$ are learnable parameters of the i-th disease representation, respectively.

Intuitively, this set-up decouples the high-dimensional image features $I \in \mathbb{R}^F$ into different low-dimensional disease space $\{d_{q,i}\}_{i=1}^{N} \in \mathbb{R}^E$, which facilitates the diversity and fluency of the follow-up generated reports, as empirically shown in Table 3.1. Meanwhile, as is also suggested by the ablation study in Table 3.6, the disease representation alone is insufficient for generating accurate medical reports. It is mainly due to a plain mingling of heterogeneous sources of information such as disease type (i.e., disease name) and disease state (e.g., positive or negative) in such a representation, which leads us to conceive a state-aware representation below.

State-aware Disease Embedding

As a remedy, a polarization module (self-attention) is incorporated to encode informative attributes by polarizing the visual features into different directions (states). For instance, these states may include "positive", "negative", "uncertain", or "unmentioned". Formally, let K be the number of states

and $\{s_j\}_{j=1}^{K} \in \mathbb{R}^E$ the state embeddings. The state-aware disease embedding $\hat{s}_i \in \mathbb{R}^E$ for the disease query $d_{q,i} \in \mathbb{R}^E$ is

$$\hat{s}_i = \sum_{j=1}^{K} \alpha_{ij} s_j, \tag{3.2}$$

where α_{ij} denotes the self-attention of $d_{q,i}$ and s_j:

$$\alpha_{ij} = \frac{\left(e^{d_{q,i}^T \cdot s_j}\right)}{\sum_{l=1}^{K} e^{\left(d_{q,i}^T \cdot s_l\right)}}. \tag{3.3}$$

Clearly, if the disease query $d_{q,i}$ has a large inner product with the state embedding s_j, s_j will be given more weight in the summation. An example is illustrated in Fig. 3.1 (Polarization via Self-attention). We called this self-attention, as the state embedding s_j is randomly initialized and is learned by maximizing the vector similarity between the disease query $d_{q,i}$ and the state embedding s_j, to minimize the multi-label classification loss.

Eddie as a Multi-label Classifier

To control how vector $d_{q,i}$ is similar to vector s_j, we treat this problem as a multi-label classification problem. In particular, let α_{ij} be the probability that the disease i-th has the state j-th (e.g., pneumonia disease $i = 5$ is positive $j = 1$). We can minimize the multi-label classification loss:

$$\mathcal{L}_C = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log(\alpha_{ij}), \tag{3.4}$$

where y_{ij} and α_{ij} are the ground-truth and predicted states (e.g., positive or negative) for the disease i-th correspondingly. For example, when $y_{ij} = 1$ and $\alpha_{ij} = 0$, the classification loss is maximum and the vector similarity between $d_{q,i}$ and s_j is very small. Therefore, the network is optimized to increase the similarity between $d_{q,i}$ and s_j such that $\alpha_{ij} \to 1$. Hence, the state-aware disease embedding $\{\hat{s}_i\}_{i=1}^{N}$ directly contains the disease state information as well as the visual information from the disease queries.

3.2.2 Report Generation

Enriched Disease Embedding

The disease embeddings $\{d_i\}_{i=1}^{N} \in \mathbb{R}^E$ (i.e., disease types / names / topics) and the state-aware disease embeddings $\{\hat{s}_i\}_{i=1}^{N}$ are entangled to form the enriched disease embeddings $X = \{x_i\}_{i=1}^{N}$,

$$x_i = d_i + \hat{s}_i, \tag{3.5}$$

with $x_i \in \mathbb{R}^E$. Hence, the enriched disease embeddings contain a rich repertoire of information including the disease topic d_i to be described (i.e., what disease), the predicted state $\hat{s}_i$ of that disease (i.e., good or bad), and the disease visual features $d_{q,i}$ (i.e., severity/details) through vector similarity of $d_{q,i}$ and s_j. A list of such enriched embedding is served as inputs to the Transformer encoder [63] for contextualization (i.e., how disease topics correlate to others),

$$\hat{x}_i = Encoder(x_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_N). \tag{3.6}$$

Report Generation Model

The Transformer decoder [63] is used to generate our medical reports, given the attended encoder sequence $\{\hat{x_i}\}_{i=1}^{N}$ and the previous ground-truth word embeddings $\{y_j\}_{j=1}^{t-1}$, by

$$\{\hat{y}_{t,k}\}_{k=1}^{S} = Decoder(y_1, y_2, ..., y_{t-1}, \hat{x}_1, \hat{x}_2, ..., \hat{x}_N), \tag{3.7}$$

where $\hat{y}_{t,k}$ is the probability of selecting the k-th word at the time-step t-th in the vocabulary set of S words. Intuitively, the Transformer model learns a mapping from the source sequence $\{x_i\}_{i=1}^{N}$ to the target medical sentences $\{y_j\}_{j=1}^{L}$, where L is the sequence length, similar to the way any language translation model learns, for example, to translate from German to English.

Finally, the loss function for the text generation task is just a cross-entropy loss over all L time-steps, as

$$\mathcal{L}_T = -\frac{1}{L}\sum_{t=1}^{L}\sum_{k=1}^{S} y_{t,k}\log(\hat{y}_{t,k}), \tag{3.8}$$

Table 3.1: An ablation study on the number of additional disease topics other than the 14 common diseases for the Open-I dataset.

Number of Topics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
14+0	0.214	0.128	0.082	0.055	0.238
14+50	0.461	**0.308**	**0.218**	0.157	0.357
14+100	**0.466**	0.307	**0.218**	**0.158**	0.358
14+200	0.439	0.295	0.210	0.153	**0.367**

where $y_{t,k}$ is the one-hot ground-truth value of selecting the k-th word. Therefore, the overall loss function for both the disease detection task and the report generation task with a balancing term $\alpha \in [0, 1]$ is

$$\mathcal{L} = \alpha \mathcal{L}_C + (1 - \alpha)\mathcal{L}_T. \tag{3.9}$$

3.3 Experiments

3.3.1 Datasets

We evaluate the disease detection and medical report generation tasks of our approach using three benchmark datasets: Chest X-ray 14 [67], Open-I [10], and Covid-19 [9]. On the Chest X-ray 14 dataset [67], for a fair comparison, we adopt the official train and test splits [1] and use 10% of the training set for validation. For the report generation task, Open-I [10] and Covid-19 [9] datasets are used to evaluate the language performance. Particularly, we use 80% of the dataset for training and validation, and the rest 20% for testing. Similarly, 10% of the training set is for validation.

Moreover, to enhance the medical report generation performance, we use Spacy [2] to extract the top-100 frequent noun phrases from each dataset as additional keywords or topics for the multi-label classification task.

[1] The chest X-ray 14 dataset with official train and test splits: https://nihcc.app.box.com/v/ChestXray-NIHCC

[2] Spacy, the industrial natural language processing tool: https://spacy.io

3.3.2 Quantitative Experiments

Disease Detection Task

In terms of the disease detection task, to be consistent with the prior works [67], [68], the commonly-used evaluation metric AUC score is adopted. As summarized in Table 3.2, both CNN-backbones obtain promising results, while our ResNeSt-50 [80] (ours-2) produces the best results. This empirical evidence suggests that the proposed polarization/state-aware mechanism can perform well in the disease detection task under different CNN backbones, which is the foundation for generating high-quality reports. Moreover, under the same ResNet-50 backbone, we compare our EDDIE method with the standard binary multi-label classification method that uses a fully connected layer (CNN+FC). Table 3.5 shows that our approach performs better than the CNN+FC.

Note that some methods have different experiment setups (e.g., non-official dataset splits [53] or training on extra datasets [19]) or require significant architecture changes [42], [43], [54] such as fine-tuning the detection networks via evolutionary algorithms. We exclude these works from our comparison and only focus on the works that share simple and widely-used CNN backbones such as ResNet-50 to avoid biases.

Medical Report Generation Task

The report generation task is evaluated on two benchmarks, Open-I and COVID-19. The widely-used language evaluation metrics are adopted, including BLEU-1 to BLEU-4 scores, ROUGE score. As for the Hit score, two board-certified radiologists are asked to evaluate the generated medical reports of 100 randomly-chosen test images. They manually mark true (1) if a generated report accurately describes the corresponding image, or mark false (0) if a generated report misses something or mistakenly predicts something. The total number of accurate reports in percentage, or Hit, is then reported. $Hit\% = \frac{\#AgreedReports}{\#TotalReports} \times 100\%$ is used to quantify the feedback of user studies.

As demonstrated in Table 3.3, our approach outperforms the compari-

Table 3.2: Disease detection performance (Ours-1: ResNet-50 and Ours-2: ResNeSt-50) on the Chest X-ray 14 benchmark.

Disease	Samples	Chest-8 [67]	TieNet [68]	Ours-1	Ours-2
Atelectasis	3,255	0.700	0.732	**0.769**	0.765
Cardiomegaly	1,065	0.810	0.844	**0.878**	0.877
Effusion	4,648	0.759	0.793	**0.825**	**0.825**
Infiltration	6,088	0.661	0.666	0.698	**0.703**
Mass	1,712	0.693	0.725	**0.812**	0.805
Nodule	1,615	0.668	0.685	0.757	**0.760**
Pneumonia	477	0.658	0.720	0.711	**0.725**
Pneumothorax	2,661	0.799	0.847	0.843	**0.852**
Consolidation	1,815	0.703	0.701	0.741	**0.745**
Edema	925	0.805	0.829	0.831	**0.834**
Emphysema	1,093	0.833	0.865	0.894	**0.909**
Fibrosis	435	0.786	0.796	0.792	**0.822**
Pleural Thickening	1,143	0.684	0.735	0.752	**0.767**
Hernia	86	0.871	**0.876**	0.853	0.862
No Finding	9,912	-	0.701	0.727	**0.731**
Average	-	0.745	0.772	0.797	**0.804**

son methods in almost all language metrics at the benchmarks (COVID-19 & Open-I). Moreover, our approach achieves the highest Hit score, manifesting its ability of generating fluent and more importantly, clinically-accurate medical reports when compared to others.

Moreover, we also disable the disease classification task and remove that state-attention module. Hence, the model only learns to generate medical reports from the implicitly learned disease representations mixed with state embedding. Table 3.6 shows that without the proposed state-aware embedding module, the performance is significantly lower than our full model. It highlights the importance of explicitly decoupling the disease features and the disease states in the ensuing the task of medical report generation.

3.4 Summary

We propose an Eddie-Transformer model to jointly tackle the tasks of robust chest X-ray disease detection and medical report generation. The proposed Eddie model decouples visual features into different states to enhance the

Table 3.3: Report generation performance with human evaluation (Hit percentage) for Open-I & COVID-19 datasets.

Dataset	Methods	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	Hit
Open-I	S&T [65]	0.224	0.136	0.095	0.072	0.300	54%
	SA&T [73]	0.256	0.156	0.109	0.081	0.303	61%
	TieNet [68]	0.330	0.194	0.124	0.081	0.311	-
	AoA [24]	0.180	0.112	0.077	0.058	0.294	51%
	CoAtt [28]	0.442	0.290	0.205	0.148	**0.387**	74%
	HRNN [75]	0.445	0.292	0.201	0.154	0.344	-
	HRGR-Agent [33]	0.438	0.298	0.208	0.151	0.322	-
	HRG-Transformer [59]	0.464	0.301	0.212	**0.158**	-	-
	Ours	**0.466**	**0.307**	**0.218**	**0.158**	0.358	**77%**
COVID-19	S&T [65]	0.179	0.082	0.040	0.020	0.155	24%
	SA&T [73]	0.159	0.074	0.038	0.020	0.155	33%
	CoAtt [28]	0.214	0.081	0.019	0.000	0.158	37%
	AoA [24]	0.055	0.027	0.014	0.008	0.099	16%
	Ours	**0.269**	**0.134**	**0.073**	**0.042**	**0.176**	**41%**

Table 3.4: Computation cost comparison. Lower is better.

Methods	MACs	Params
Ours (Eddie-Transformer)	**230.2 billion**	**56 million**
CoAtt(Hierarchical LSTM)	582.5 billion	342.6 million

Table 3.5: Comparison of the traditional multi-label classifier and our approach on the Chest X-ray 14 dataset (the AUC score is the average of all 14 common diseases).

Detection	CNN+FC	Ours
No Finding	0.714	**0.727**
Avg Finding	0.778	**0.797**

Table 3.6: The ablation study of our approach w/ vs. w/o the EDDIE module, evaluated on the Open-I dataset.

Methods	B-1	B-2	B-3	B-4	ROUGE
No EDDIE	0.399	0.254	0.177	0.131	0.338
W/ EDDIE	**0.466**	**0.307**	**0.218**	**0.158**	**0.358**

visual feature representations and increase the report's quality. Besides, the disease-state aware mechanism can also be used as a disease detection model. Our empirical evaluations demonstrate our approach's effectiveness on disease detection and medical report generation tasks on three different benchmarks: Chest X-ray 14, Open-I, and COVID-19. Thanks to the good performance, our proposed model plays as the core framework for further improvements in the next chapter.

Chapter 4

Automated Generation of Accurate & Fluent Medical X-ray Reports

4.1 Introduction

A successful medical report generation process is expected to possess two key properties:

- Clinical accuracy: properly and correctly describing the disease and related symptoms.
- Language fluency: producing realistic and human-readable text.

Many recent progresses in the medical report generation [27], [28], [33], [34], [40], [59], [68], [74], [75], [78] often perform reasonably well in addressing the language fluency aspect. On the other hand, as is also evidenced in our empirical evaluation, their results are notably less satisfactory in terms of clinical accuracy.

This we attribute to several reasons: one is closely tied to the textual characteristic of medical reports, which typically consists of many long sentences describing various disease-related symptoms and related topics in precise and domain-specific terms. Moreover, the disease patterns are not semantic for medical report generators to generate meaningful descriptions. Another reason is related to the lack of full use of rich contextual information that encodes prior knowledge. This information includes, for example, the patient's clinical

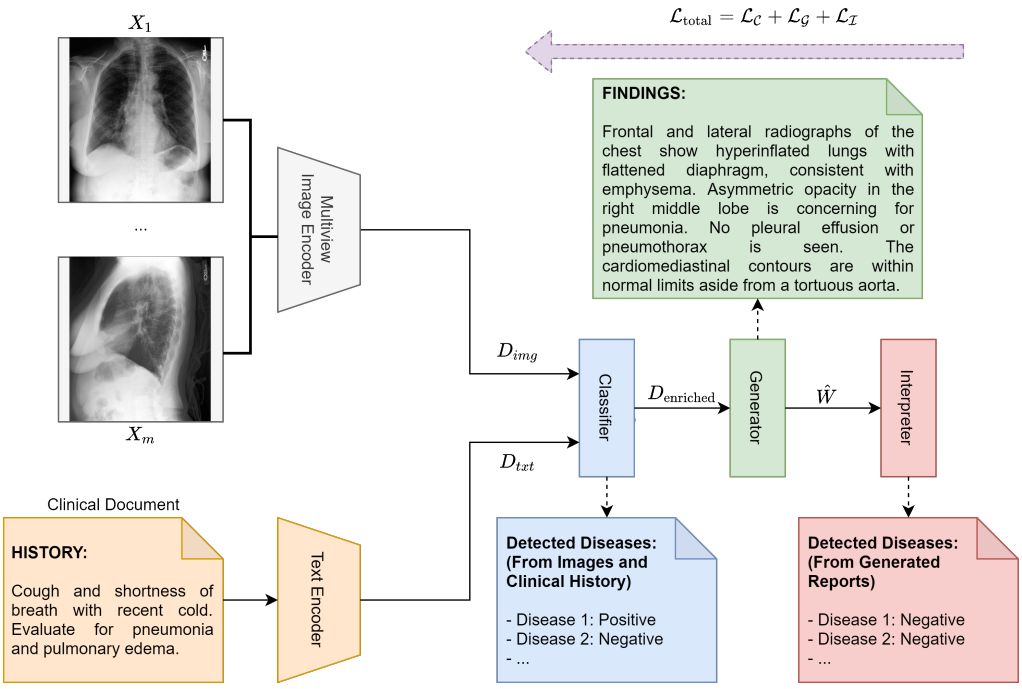


Figure 4.1: Our approach consists of three modules: a *classifier* that reads chest X-ray images and clinical history to produce an internal checklist of disease-related topics, a transformer-based *generator* to generate fluent text, and an *interpreter* to examine and fine-tune the generated text to be consistent with the disease-related topics.

document describing the key clinical history and indication from doctors, and multiple scans from distinct 3D views – information that is typically existed in abundance in practical scenarios, as, e.g., in the standard X-ray benchmarks of Open-I [10] and MIMIC-CXR [29]. Last but not least, the report generation module may cheat the learning process, ignore the outputs of the classification task, and bias towards maximizing language evaluation metrics.

For these reasons, we propose a categorize-generate-interpret framework that places specific emphasis on clinical accuracy while maintaining adequate language fluency of the generated reports. It consists of a classifier module that reads chest X-ray images (e.g., either single-view or multi-view images) and related documents to detect diseases and output enriched disease embedding; a transformer-based medical report generator to robustly generate long paragraphs; and a differentiable interpreter to evaluate and fine-tune the generated reports for factual correctness. The main contributions are two-fold:

- A differentiable end-to-end approach is proposed, consisting of three modules (classifier-generator-interpreter): the classifier module learns the disease feature representation via context modeling (section 4.2.1) and disease-state aware mechanism (section 4.2.1); the generator module transforms the disease embedding to medical report; the interpreter module reads and fine-tunes the generated reports, enhancing the consistency of the generated reports and the classifier's outputs.

- Empirically, our approach is shown to outperform against many strong baselines over two widely-used benchmarks on an equal footing (i.e., without access to additional information). Moreover, empirical evidence demonstrates that clinical patient history and additional scans may play a vital role in improving the quality of the generated reports.

4.2 Our Approach

Our framework consists of a *classification* module, a *generation* module, and an *interpretation* module, as illustrated in Fig. 4.1. The classification module reads multiple chest X-ray images and extracts the global visual feature representation via a multi-view image encoder. They are then disentangled into multiple low-dimensional visual embedding. Meanwhile, the text encoder reads clinical documents, including, e.g., doctor indication, and summarizes the content into text-summarized embedding. The visual and text-summarized embeddings are entangled via an "add & layerNorm" operation to form contextualized embedding in terms of disease-related topics. The generation module takes our enriched disease embedding as initial input and generates text word-by-word, as shown in Fig. 4.2. Finally, the generated text is fed to the interpretation module for fine-tuning to align to the checklist of disease-related topics from the classification module. In what follows, we are to elaborate on these three modules in detail.

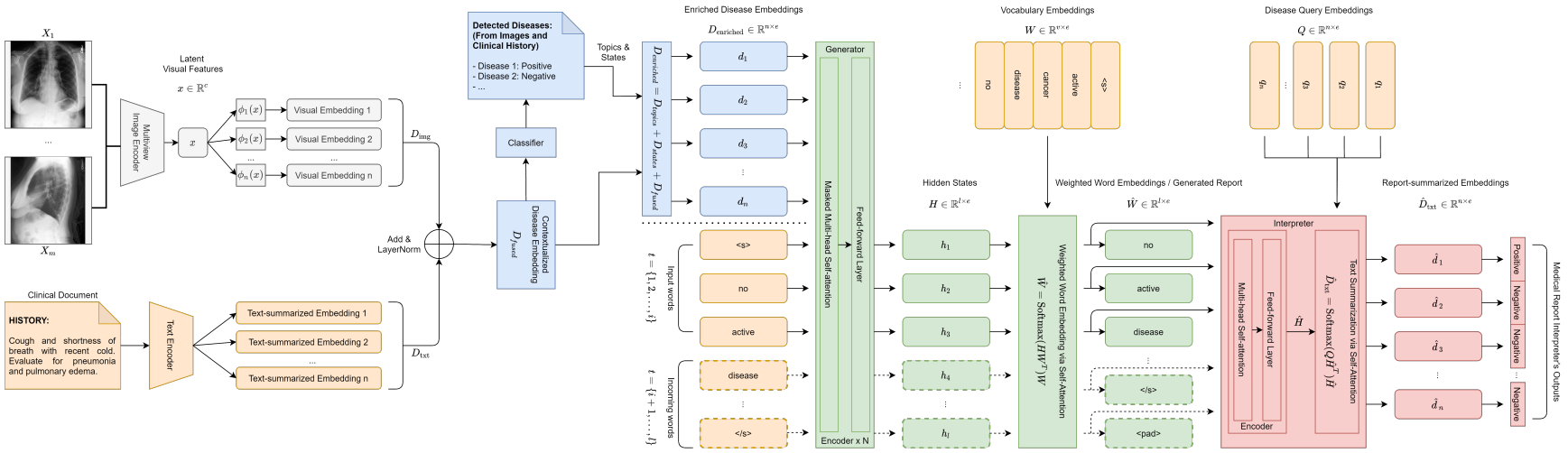


Figure 4.2: An exemplar illustration of our approach in action. Specifically, the enriched disease embedding produced from the classification module are fed into the generation module as initial inputs. Then, at each time step, the hidden state h_i is obtained to predict the next output word. Finally, the interpretation module takes as input all predicted outputs $\hat{W}$ to produce a checklist of disease-related topics, which are to be gauged with the same topics output from the classification module for consistency verification.

4.2.1 The Classification Module

Multi-view Image Encoder

For each medical study which consists of m chest X-ray images $\{X_i\}_{i=1}^m$, we extract the corresponding latent features $\{x_i\}_{i=1}^m \in \mathbb{R}^c$, where c is the number of features, via a shared DenseNet-121 image encoder [22]. Then, the multi-view latent features $x \in \mathbb{R}^c$ can be obtained by max-pooling across the set of m latent features $\{x_i\}_{i=1}^m$, as proposed in [60]. When $m = 1$, the multi-view encoder boils down to a single-image encoder.

Text Encoder

Let T be a text document with length l consisting of word embeddings $\{w_1, w_2, ..., w_l\}$, where $w_i \in \mathbb{R}^e$ embodies the i-th word in the text and e is the embedding dimension. We use the transformer encoder [63] as our text feature extractor to retrieve a set of hidden states $H = \{h_1, h_2, ..., h_l\}$, where $h_i \in \mathbb{R}^e$ is the attended features of the i-th word to other words in the text,

$$h_i = \text{Encoder}(w_i | w_1, w_2, ..., w_l). \tag{4.1}$$

The entire document T is then summarized by $Q = \{q_1, q_2, ..., q_n\}$, representing n disease-related topics (e.g., pneumonia or atelectasis) to be queried from the document. We refer to this retrieval process as *text-summarized embedding* $D_{\text{txt}} \in \mathbb{R}^{n \times e}$,

$$D_{\text{txt}} = \text{Softmax}\left(QH^\intercal\right) H. \tag{4.2}$$

Here matrix $Q \in \mathbb{R}^{n \times e}$ is formed by stacking the set of vectors $\{q_1, q_2, ..., q_n\}$ where $q_i \in \mathbb{R}^e$ is randomly initialized, then learned via the attention process. Similarly, the matrix $H \in \mathbb{R}^{l \times e}$ is formed by $\{h_1, h_2, ..., h_l\}$ from Eq. (4.1). The term $\text{Softmax}(QH^\intercal)$ is the word attention heat-map for the n queried diseases in the document. The intuition here is for each disease (e.g., pneumonia) to be queried from the text document T. We only pay attention to the most relevant words (e.g., *cough* or *shortness of breath*) in the text that associates with that disease, also known as a vector similarity dot product. This way, the

weighted sum of these words by Eq. (4.2) gives the feature that summarizes the document w.r.t. the queried disease.

Contextualized Disease Embedding

The latent visual features $x \in \mathbb{R}^c$ are subsequently decoupled into low-dimensional disease representations, as illustrated in Fig. 4.1. They are regarded as the *visual embedding* $D_{\text{img}} \in \mathbb{R}^{n \times e}$, where each row is a vector $\phi_j(x) \in \mathbb{R}^e, j = 1, \ldots, n$ defined as follows:

$$\phi_j(x) = A_j^\intercal x + b_j. \tag{4.3}$$

Here $A_j \in \mathbb{R}^{c \times e}$ and $b_j \in \mathbb{R}^e$ are learnable parameters of the j-th disease representation. n is the number of disease representations, and e is the embedding dimension. Now, together with the available clinical documents, the visual embedding D_{img} and the text-summarized embedding D_{txt} are entangled to form *contextualized disease representations* $D_{\text{fused}} \in \mathbb{R}^{n \times e}$ as

$$D_{\text{fused}} = \text{LayerNorm}(D_{\text{img}} + D_{\text{txt}}). \tag{4.4}$$

Intuitively, the entanglement of visual and textual information allows our model to mimic the hospital workflow, to screen the disease's visual representations conditioned on the patients' clinical history or doctors' indication. For example, the doctor's indication in Fig. 4.1 shows *cough* and *shortness of breath* symptoms. It is reasonable for a medical doctor to request a follow-up check of the *pneumonia* disease. As for the radiologists receiving the doctors' indication, they may prioritize diagnosing the presence of *pneumonia* and related diseases based on X-ray scans and look for specific abnormalities. As empirically shown in Table 4.5, the proposed contextualized disease representations bring a significant performance boost in the medical report generation task. Meanwhile, our current embedding is basically a plain mingling of heterogeneous sources of information such as disease type (i.e., disease name) and disease state (e.g., positive or negative). As shown by the ablation study in Table 4.5, this embedding by itself is insufficient for generating accurate medical reports. This leads us to conceive a follow-up enriched representation below.

Enriched Disease Embedding

The main idea behind *enriched disease embedding* is to further encode informative attributes about disease states, such as *positive*, *negative*, *uncertain*, or *unmentioned*. Formally, let k be the number of states and $S \in \mathbb{R}^{k \times e}$ the state embedding. Then the confidence of classifying each disease into one of the k disease states is

$$p = \text{Softmax}(D_{\text{fused}} S^{\intercal}). \tag{4.5}$$

$S \in \mathbb{R}^{k \times e}$ is randomly initialized, then learned via the classification of D_{fused}. D_{fused} acts as features for the multi-label classification, and the classification loss is computed as

$$\mathcal{L}_{\mathcal{C}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \log(p_{ij}), \tag{4.6}$$

where $y_{ij} \in \{0, 1\}$ and $p_{ij} \in (0, 1)$ are the j-th ground-truth and predicted values for the disease i-th, respectively. The state-aware embedding $D_{\text{states}} \in \mathbb{R}^{n \times e}$ are then computed as

$$D_{\text{states}} = \begin{cases} yS, & \text{if training phase} \\ pS, & \text{otherwise.} \end{cases} \tag{4.7}$$

$y \in \{0, 1\}^{n \times k}$ is the one-hot ground-truth labels about the disease-related topics, whereas $p \in (0, 1)^{n \times k}$ is the predicted values. During training, the ground-truth disease states facilitate our generator in describing the diseases & related symptoms based on accurate information (teacher forcing). At test time, our generator then furnishes its recount based on the predicted states.

Finally, the *enriched disease embedding* $D_{\text{enriched}} \in \mathbb{R}^{n \times e}$ is the composition of state-aware disease embedding D_{states} (i.e., good or bad), disease names D_{topics} (i.e., which disease/topic), and the disease representations D_{fused} (i.e., severity and details of the diseases),

$$D_{\text{enriched}} = D_{\text{states}} + D_{\text{topics}} + D_{\text{fused}}. \tag{4.8}$$

Like the disease queries Q, $D_{\text{topics}} \in \mathbb{R}^{n \times e}$ is randomly initialized, representing diseases or topics to be generated. It is then learned in training through the medical report generation pipeline. The enriched disease embedding provides

explicit and precise disease descriptions, and endows our follow-up generation module with a powerful data representation.

4.2.2 The Generation Module

Our report generator is derived from the transformer encoder of [63]. The network is formed by sandwiching & stacking a masked multi-head self-attention component and a feed-forward layer being on top of each other for N times, as illustrated in Fig. 4.2. The hidden state for each word position $h_i \in \mathbb{R}^e$ in the medical report is then computed based on previous words and disease embedding, as $D_{\text{enriched}} = \{d_i\}_{i=1}^n$,

$$h_i = \text{Encoder}(w_i | w_1, w_2, ..., w_{i-1}, d_1, d_2, ..., d_n). \tag{4.9}$$

This is followed by predicting future words based on the hidden states $H = \{h_i\}_{i=1}^l \in \mathbb{R}^{l \times e}$, as

$$p_{\text{word}} = \text{Softmax}(HW^\intercal). \tag{4.10}$$

Here $W \in \mathbb{R}^{v \times e}$ is the entire vocabulary embedding, v the vocabulary size, and l the document length. Let $p_{\text{word},ij}$ denote the confidence of selecting the j-th word in the vocabulary W for the i-th position in the generated medical report. The generator loss is defined as a cross entropy of the ground-truth words y_{word} and predicted words p_{word},

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{v} y_{\text{word},ij} \log(p_{\text{word},ij}). \tag{4.11}$$

Finally, the *weighted word embedding* $\hat{W} \in \mathbb{R}^{l \times e}$, also known as the *generated report*, are:

$$\hat{W} = p_{\text{word}} W. \tag{4.12}$$

It is worth noting that this set-up facilitate the back-propagation of errors from the follow-up interpretation module.

4.2.3 The Interpretation Module

It is observed from empirical evaluations that the generated reports are often distorted in the process, such that they become inconsistent with the original

output of the classification module – the enriched disease embedding that encodes the disease and symptom related topics. Inspired by the CycleGAN idea of [82], we consider a fully differentiable network module to estimate the checklist of disease-related topics based on the generator's output, and to compare with the original output of the classification module. This provides a meaningful feedback loop to regulate the generated reports, which is used to fine-tune the generated report through the word representation outputs $\hat{W}$.

Specifically, we build on top of the proposed text encoder (described in section 4.2.1) a classification network that classifies disease-related topics, as follows. First, the text encoder summarizes the current medical report $\hat{W}$, and outputs the report-summarized embedding of the queried diseases Q,

$$\hat{D}_{\text{txt}} = \text{Softmax}(Q\hat{H}^{\intercal})\hat{H} \in \mathbb{R}^{n\times e}. \tag{4.13}$$

Here $\hat{H}$ is computed from the generated medical reports $\hat{W}$ using Eq. (4.1). Second, each of the report-summarized embedding $\hat{d}_i \in \mathbb{R}^e$ (i.e., each row of the matrix $\hat{D}_{\text{txt}} \in \mathbb{R}^{n\times e}$) is classified into one of the k disease-related states (i.e., positive or negative), as

$$p_{\text{int}} = \text{Softmax}(\hat{D}_{\text{txt}}S^{\intercal}) \in \mathbb{R}^{n\times k}. \tag{4.14}$$

Finally, the interpreter is trained to minimize the subsequent multi-label classification loss,

$$\mathcal{L}_{\mathcal{I}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log(p_{\text{int},ij}). \tag{4.15}$$

here $y_{ij} \in \{0,1\}$ is the ground-truth disease label and $p_{\text{int},ij} \in (0,1)$ is the predicted disease label of the interpreter.

In fine-tuning the generated medical reports $\hat{W}$, all interpreter parameters are frozen, which acts as a guide to force the word representations $\hat{W}$ being close to what the interpreter has learned from the ground-truth medical reports. If the weighted word embedding $\hat{W}$ is different from the learned representation – which leads to incorrect classification – a large loss value will be imposed in the interpretation module. This thus forces the generator to move toward producing a correct word representation.

Datasets	Methods	B-1	B-2	B-3	B-4	MTR	RG-L	SV	MV	AI	FT
Open-I	S&T [65]	0.316	0.211	0.140	0.095	0.159	0.267	x			
	LRCN [13]	0.369	0.229	0.149	0.099	0.155	0.278	x			
	SA&T [73]	0.399	0.251	0.168	0.118	0.167	0.323	x			
	Att-RK [77]	0.369	0.226	0.151	0.108	0.171	0.323	x			
	HRNN [76]	0.445	0.292	0.201	0.154	0.175	0.344	x			
	1-NN [3]	0.232	0.116	0.051	0.018	N/A	0.201	x			
	TieNet [68]	0.330	0.194	0.124	0.081	N/A	0.311	x			
	Liu et. al. [38]	0.359	0.237	0.164	0.113	N/A	0.354	x			x
	CoAtt [28]	0.455	0.288	0.205	0.154	N/A	0.369	x			
	HRGR-Agent [33]	0.438	0.298	0.208	0.151	N/A	0.322		x		x
	KERP [34]	0.482	0.325	0.226	0.162	N/A	0.339		x	x	
	ReinforcedTransformer [72]	0.350	0.234	0.143	0.096	N/A	N/A	x			x
	HRG-Transformer [59]	0.464	0.301	0.212	0.158	N/A	N/A		x		
	SD&C [27]	0.464	0.301	0.210	0.154	N/A	0.362	x			x
	Ours (SV)	0.463	0.310	0.215	0.151	0.186	0.377	x			
	Ours (MV)	0.476	0.324	0.228	0.164	0.192	0.379		x		
	Ours (MV+T)	0.485	0.355	0.273	0.217	0.205	0.422		x	x	
	Ours (MV+T+I)	**0.515**	**0.378**	**0.293**	**0.235**	**0.219**	**0.436**		x	x	x
MIMIC	1-NN [3]	0.367	0.215	0.138	0.095	0.139	0.228	x			
	SA&T [73]	0.370	0.240	0.170	0.128	0.141	0.310	x			
	AdpAtt [41]	0.384	0.251	0.178	0.134	0.148	0.314	x			
	Liu et. al. [38]	0.313	0.206	0.146	0.103	N/A	0.306	x			x
	Transformer [63]	0.409	0.268	0.191	0.144	0.157	0.318	x			
	GumbelTransformer [40]	0.415	0.272	0.193	0.146	0.159	0.318	x			x
	Ours (SV)	0.447	0.290	0.200	0.144	0.186	0.317	x			
	Ours (MV)	0.451	0.292	0.201	0.144	0.185	0.320		x		
	Ours (MV+T)	0.491	0.357	0.276	0.223	0.213	0.389		x	x	
	Ours (MV+T+I)	**0.495**	**0.360**	**0.278**	**0.224**	**0.222**	**0.390**		x	x	x

Table 4.1: Quantitative comparison of our approach and a number of recent works. Since these works are evaluated under different setups of Single-view (SV), Multi-view (MV), w/ clinical text (T), and interpreter (I), for a fair comparison, all methods are categorized based on the following four aspects: Single-View (SV), Multi-view (MV), Additional Information (AI), and Fine-tuning of the generated reports (FT). Best results are highlighted in **bold face**. Different language metrics are employed, including BLEU-1 to BLEU-4 (B-1 to B-4), METEOR (MTR), and ROUGE-L (RG-L).

Collectively our model is trained in an end-to-end manner by jointly minimizing the total loss,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{I}}. \tag{4.16}$$

4.3 Experiments

This section evaluates the medical report generation task on two fronts: the language performance and the clinical accuracy performance. Empirical evaluations are carried out on two widely-used chest X-ray datasets, MIMIC-CXR [29] and Open-I [10].

4.3.1 Datasets

MIMIC-CXR Dataset

The MIMIC-CXR dataset [29] is a large-scale dataset with 227,835 medical reports of 65,379 patients, associated with 377,110 images from multiple views:

anterior-posterior (AP), posterior-anterior (PA), lateral (LA). Each study comprises multiple sections, including *comparison*, *clinical history*, *indication*, *reasons for examination*, *impressions*, and *findings*. Here we utilize the multi-view images of AP/PA/LA views, and adopt as contextual information the concatenation of the *clinical history*, *reason for examination*, and *indication* sections. For consistency, we follow the experimental set-up of [40] to focus on generating text in the "findings" section as the corresponding medical report.

Open-I Dataset

The Open-I dataset [10] collected by the Indiana University hospital network contains 3,955 radiology studies that correspond to 7,470 frontal and lateral chest X-rays. Some radiology studies are associated with more than one chest X-ray image. Each study typically consists of *impression*, *findings*, *comparison*, and *indication* sections. Similar to the MIMIC-CXR dataset, we utilized both the multi-view chest X-ray images (frontal and lateral) and the *indication* section as our contextual inputs. For generating medical reports, we follow the existing literature [28], [59] by concatenating the *impression* and the *findings* sections as the target output.

An important note: the implementation details, dataset splits, preprocessing steps, generated examples, and qualitative analysis are described in the supplementary materials.

4.3.2 Experimental Results

Language Generation Performance

A comprehensive quantitative comparison of our approach and many baselines as shown in Table 4.1 on the two benchmarks using the widely-used language evaluation metrics: BLEU-1 to BLEU-4 [50], ROUGE-L [36], and METEOR [1] scores. Since all comparison methods have their own experiment setups, for a fair comparison, we further categorize these methods into four aspects: single-view (SV), multi-view (MV), accessing to additional information (AI) such as clinical document, and applying fine-tuning (FT) to the

Symbol	Meaning	Size
n	Num. of disease-related topics	114
e	Embedding dimension	256
c	Num. of visual features	1024
k	Num. of states	2
l	Max document length	1000
m	Num. of multi-view images	2
v	Vocabulary size	1000
$\mathbf{w_i}$	Word embedding	$\mathbb{R}^e$
$\mathbf{h_i}$	Attended features	$\mathbb{R}^e$
$\mathbf{s_i}$	The i-th state embedding	$\mathbb{R}^e$
$\phi_\mathbf{i}(\mathbf{x})$	Visual transformation	$\mathbb{R}^e$
$\mathbf{x_i}$	Visual features of the i-th view	$\mathbb{R}^c$
$\mathbf{x}$	Multi-view visual features	$\mathbb{R}^c$
S	State embedding	$\mathbb{R}^{k\times e}$
H	Hidden states (attended features)	$\mathbb{R}^{l\times e}$
$\hat{H}$	Hidden states of gen. reports	$\mathbb{R}^{l\times e}$
W	Vocabulary embedding	$\mathbb{R}^{v\times e}$
$\hat{W}$	Weighted word embedding	$\mathbb{R}^{l\times e}$
T	The input text document	$\mathbb{R}^{l\times e}$
X_i	View i-th chest X-ray image	$\mathbb{R}^{256\times 256}$
Q	Disease query embedding	$\mathbb{R}^{n\times e}$
D_{img}	Visual embedding	$\mathbb{R}^{n\times e}$
D_{txt}	Text-summarized embedding	$\mathbb{R}^{n\times e}$
$\hat{D}_{\text{txt}}$	Report-summarized embedding	$\mathbb{R}^{n\times e}$
D_{fused}	Contextualized disease emb.	$\mathbb{R}^{n\times e}$
D_{topics}	Topic embedding	$\mathbb{R}^{n\times e}$
D_{states}	State-aware embedding	$\mathbb{R}^{n\times e}$
D_{enriched}	Enriched disease embedding	$\mathbb{R}^{n\times e}$
p_{ij}	Predicted outputs	$(0, 1)$
y_{ij}	One-hot ground-truth outputs	$\{0, 1\}$
$\mathcal{L}$	Loss functions	$\mathbb{R}$

Table 4.2: The summary of our notation and symbols.

Datasets	Methods	Acc.	Macro scores				Micro scores			
			AUC	F-1	Prec.	Rec.	AUC	F-1	Prec.	Rec.
Open-I	1-NN [3]	0.911	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	S&T [65]	0.915	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	SA&T [73]	0.908	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	TieNet [68]	0.902	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Liu et. al. [38]	0.918	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Ours (SV)	0.944	0.595	0.118	0.125	0.136	0.857	0.657	0.651	**0.663**
	Ours (MV)	0.943	0.626	0.144	0.149	0.150	**0.878**	0.648	0.647	0.649
	Ours (MV+T)	**0.947**	0.671	0.130	**0.192**	0.124	0.873	**0.659**	**0.687**	0.634
	Ours (MV+T+I)	0.937	**0.702**	**0.152**	0.142	**0.173**	0.877	0.626	0.604	0.649
MIMIC	1-NN [3]	N/A	N/A	0.206	0.213	0.200	N/A	0.335	0.346	0.324
	SA&T [73]	N/A	N/A	0.101	0.247	0.119	N/A	0.282	0.364	0.230
	AdpAtt [41]	N/A	N/A	0.163	0.341	0.166	N/A	0.347	0.417	0.298
	Liu et. al. [38]	0.867	N/A	N/A	0.309	0.134	N/A	N/A	**0.586**	0.237
	Transformer [63]	N/A	N/A	0.214	0.327	0.204	N/A	0.398	0.461	0.350
	GumbelTransformer [40]	N/A	N/A	0.228	0.333	0.217	N/A	0.411	0.475	0.361
	Ours (SV)	0.877	0.743	0.342	0.357	0.347	0.857	0.530	0.533	0.528
	Ours (MV)	0.880	0.752	0.347	0.385	0.347	0.862	0.533	0.545	0.522
	Ours (MV+T)	**0.890**	0.778	0.407	**0.448**	0.399	0.872	**0.578**	0.583	0.574
	Ours (MV+T+I)	0.887	**0.784**	**0.412**	0.432	**0.418**	**0.874**	0.576	0.567	**0.585**

Table 4.3: Quantitative comparison of clinical accuracy from the generated reports of a number of recent methods, evaluated on the 14 common CheXpert's diseases. The best results are highlighted in **bold face**.

generated medical reports. Experiments in Table 4.1 show that our models outperform the baselines in most language metrics.

With a single input X-ray image as the sole input, ours (SV) outperforms by a noticeable margin the best SOTA methods of CoAtt on Open-I and Transformer on MIMIC, respectively. This we mainly attribute to the utilization of the enriched disease embedding that explicitly incorporates the disease-related topics. With multiple X-ray images as input, Ours (MV) again outperforms the best comparison methods of HRG-Transformer on Open-I. With multiple X-ray images and additional clinical document information as input, ours (MV+T) outperforms the comparison methods of KERP on Open-I. Finally, with the complete contextual information available as input, ours (MV+T+I) outperforms all the comparison methods available in both Open-I and MIMIC datasets.

Clinical Accuracy Performance

To evaluate the clinical accuracy of the generated reports, we use the LSTM CheXpert labeler [40] as a universal measurement. We compare different methods based on accuracy, F-1, precision (prec.), and recall (rec.) metrics on 14 common diseases. Since there are 14 independent diseases, we also report the macro and micro scores. Intuitively, a high macro score means the detection of all 14 diseases is improved. Meanwhile, a high micro score implies the

dominant diseases are improved (i.e., some diseases appear more frequently than others). As observed in Table 4.3, our clinical performance increased significantly compared to the baselines in both macro and micro scores.

Among our ablation models in Table 4.3, the precision and accuracy scores of our contextualized variant (MV+T) tend to be higher, whereas other scores are lower than the one with the interpreter (MV+T+I). This opposite behavior is due to the interpreter, which encourages detecting diseases, thus increases False Positives (FP). Note in the medical context, it is usually critically important to lower the False Negatives (FN) rate, thus a high recall score with a slight decrease in precision is more preferred.

Human Evaluation

In addition to the automated evaluations, we ask an experienced medical doctor to evaluate our generated medical reports. Specifically, the chest X-ray images and ground-truth medical reports are given to the doctor. Then, the doctor evaluates the quality of the generated reports by assigning a score from 0 (totally disagree) to 10 (totally agree). The final score for each model is computed by averaging all scores (97 test samples for each proposed model).

It can be inferred from Table. 4.4 that the MV+T+I gives more accurate medical reports and using the interpreter to fine-tune the outputs is indeed improving the reports' quality. Additionally, it is also clear from the human evaluation that incorporating clinical history information positively affects the final performance. Moreover, the human evaluation shows that most generated examples are good (8.031 on average), indicating the proposed model's effectiveness in terms of clinical accuracy.

Qualitative Analysis

Figure. 4.3 showcases the generated examples when engaging our full-fledged approach. It is clear that our approach is capable of generating closely matched descriptions for both healthy cases (the first 3 examples) and disease cases (the last 3 examples). From the last 3 examples, our generated reports correctly detect diseases including *pleural effusions*, *atelectasis*, as well as surgical and

Methods	Average	Min	Max	Median	Q1	Q3
Ours (SV)	7.000	**3**	**10**	7	6	8
Ours (MV)	7.237	**3**	**10**	7	**7**	8
Ours (MV+T)	7.794	**3**	**10**	**8**	**7**	**9**
Ours (MV+T+I)	**8.031**	**3**	**10**	**8**	**7**	**9**

Table 4.4: The human evaluation scores for the generated reports from an experienced medical doctor. For each model, we take the average, min, max, median, first, and the third quartile of all ratings given by the doctor. The score is in the range of 0 (totally disagree) to 10 (totally agree).

supporting devices such as *wires* and *clips*.

In terms of failure cases, our results still contain False Positive cases at times: one is in detecting *atelectasis* in the last example, which could not match up with anything in the ground-truth report; In the second last example, our report confuses between *left* and *right* atelectasis. This is because our extracted visual features do not explicitly account for orientation and direction. Similarly, our interpreter is only used to promote disease detection (i.e., the presence or absence of diseases).

Previously, we mentioned that the proposed model with the interpreter (MV+T+I) tends to produce more accurate and fluent reports than the model without the interpreter (MV+T). In this part, we provide some generated examples to support our claim. We use the same six examples shown in Figure. 4.3; the chest X-ray images are omitted for clarity. As can be seen in Figure. 4.4, the contents produced from the interpreter model are more similar to the ground-truth reports than without having the interpreter. For instance, in the second last example, we highlighted some sentences about "tubes" which do not mention in the ground-truth reports. Moreover, the text does not mention "pulmonary edema" disease. In the last example, it can be seen that without the interpreter, the generated report is missing some topics such as "right pleural effusion" or "lobe opacity". These examples lead us to believe that the interpreter is indeed making the generated reports more accurate or "on-point" than the conventional "image-to-text" models.

Methods	B-1	B-2	B-3	B-4	MTR	RG-L
R w/o D_{states}	0.400	0.253	0.175	0.127	0.166	0.362
R w/o D_{topics}	0.453	0.300	0.206	0.142	0.183	0.366
R w/o D_{fused}	0.468	0.310	0.215	0.151	0.189	0.373
R with D_{enriched}	0.463	0.310	0.215	0.151	0.186	0.377
R + Interpreter	0.470	0.314	0.220	0.158	0.192	0.375
C w/o D_{states}	0.404	0.286	0.215	0.169	0.183	0.396
C w/o D_{topics}	0.474	0.329	0.244	0.187	0.194	0.401
C w/o D_{fused}	0.470	0.337	0.257	0.204	0.212	0.408
C with D_{enriched}	0.485	0.355	0.273	0.217	0.205	0.422
C + Interpreter	**0.515**	**0.378**	**0.293**	**0.235**	**0.219**	**0.436**

Table 4.5: The table compares a regular image-to-text version (R) and a contextualized version (C) of our proposed method that utilizes clinical history on the Open-I dataset. For each version, we evaluate the importance of each component D_{states}, D_{topics}, and D_{fused} in the proposed enriched disease embedding D_{enriched} by removing one component at a time.

4.3.3 Ablation studies

Enriched disease embedding

We observe that the latent features D_{fused} extracted from the classifier are insufficient to generate robust medical reports, as shown in Table 4.5. Based on our human languages, a meaningful story needs three factors: the topic (i.e., what disease), the tone (i.e., is it negative or positive), and the details (i.e., the severity). However, there is no guarantee that the learned latent features D_{fused} has all three required elements. On the other hand, with the the explicit representations (i.e., D_{fused}, D_{topics}, and D_{states}), all three factors are preserved. Therefore, the enriched disease embedding D_{enriched} can generate precise and complete medical reports, leading to the language metrics' substantial improvement.

Contextualized embedding

Table 4.5 also shows that our proposed "contextualized" version can improve the language scores over the "regular" version, which reads only images. Notably, the contextualized version is the entanglement of the chest X-ray images and the clinical history, which is crucial to improve the generated report's quality and accommodate doctors' practical needs. It mimics how radiologists

receive requests from medical doctors and write reports to answer their questions. Hence, the generated reports are believed to be more "on point" and receives higher language scores than the regular "image-to-text" setting.

4.4 Limitations and Future Work

Our work has several limitations that future works can take into consideration for further improvement. Firstly, our model does not explicitly consider disease orientation or direction (e.g., left or right, top or bottom). For example, future works can include visual-semantic embedding (direction/orientation/location) to learn and localize diseases during generating medical reports. Secondly, our work does not support time-series relationships between different studies of a patient. This information is vital to analyze existing diseases by comparing their size or structure to determine if a disease is getting worse or not. If these limitations can be addressed, the medical report system can be much more reliable for real-world applications.

Noticeably, we observe some hallucination facts (False Positives) where some diseases are mistakenly described as positive in the medical reports. For example, some images with "pneumonia" are wrongly described as "pulmonary edema". In fact, human radiologists often mistakenly classify some diseases [56]. For example, [56] shows that human radiologists or physicians can accurately detect normal lung X-ray images almost all of the time; but, for abnormal lung X-ray images, the correctness of diagnosis drops to only 50% [56]. For this reason, it is challenging to generate accurate medical reports even for experienced radiologists.

In the future, we will expand our work to related medical applications such as retinal and brain medical report generation on X-ray/MRI/CT scans. We believe that our model can be generalized to a wide range of medical report generation problems where common symptoms or disease labels and medical reports are available in most medical scan datasets. Moreover, extending the current work to incorporate tabular data inputs could be another exciting direction because some clinical information is in the form of tabular struc-

ture such as patient's age, heart pressure, or temperature [9]. In some cases, physicians must include this information in medical reports, which cannot be inferred from only reading medical scans.

4.5 Summary

This section introduces a novel three-module approach for generating medical reports from X-ray scans. Superior performance of our approach over state-of-the-art methods has been empirically demonstrated on widely-used benchmarks with a range of evaluation metrics. Our approach is also flexible and can work with additional input information, where consistent performance gains are observed.

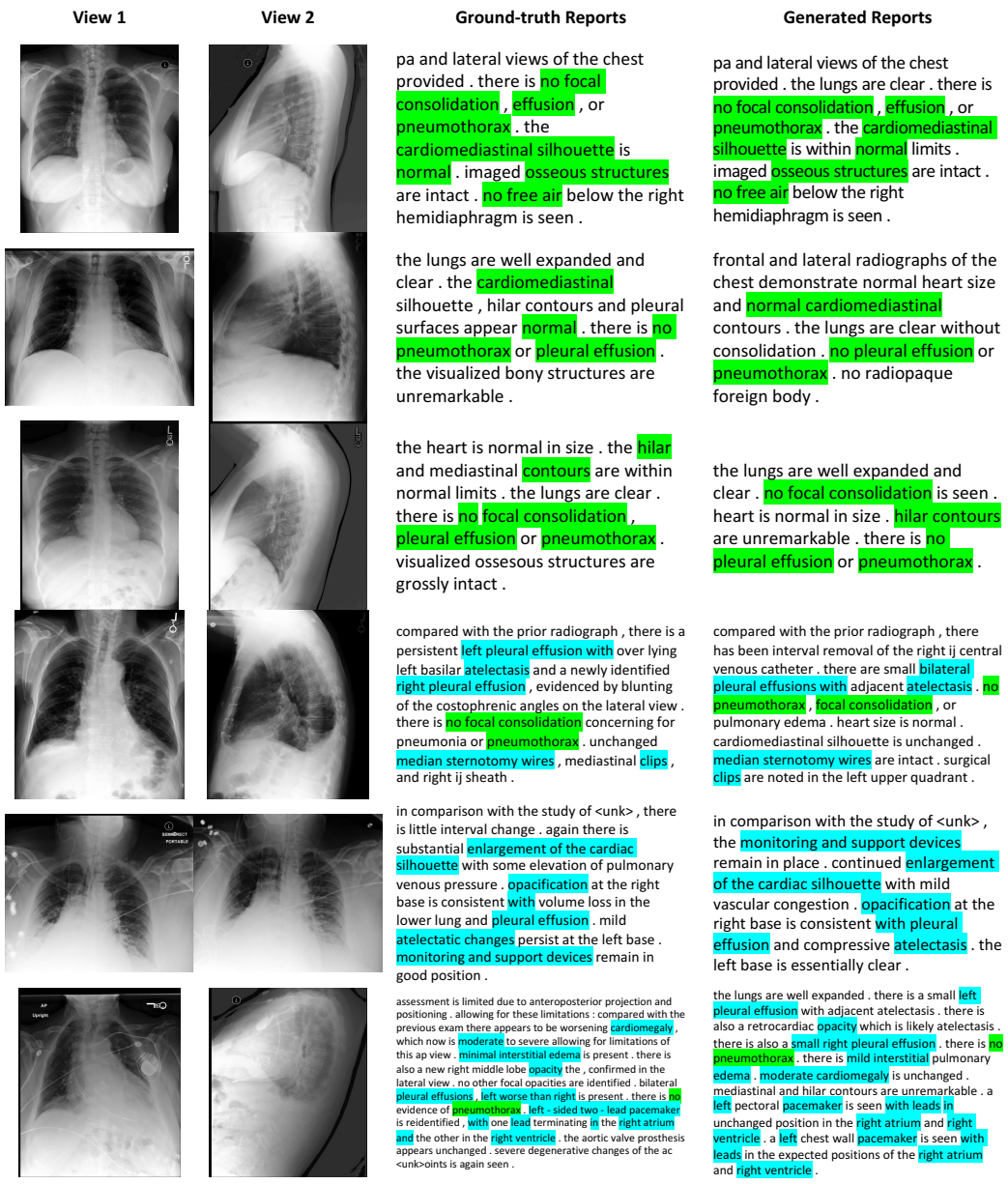


Figure 4.3: Exemplar generated medical reports of our approach for *normal* and disease cases. The matched *normal* phrases are highlighted in green-color, whereas the cyan-colored phrases are for matched diseases.

Ground-truth Reports	w/o Interpreter (MV+T)	w/ Interpreter (MV+T+I)
pa and lateral views of the chest provided . there is no focal consolidation , effusion , or pneumothorax . the cardiomediastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .	the lungs are well - expanded and clear . no focal consolidation . the cardiomediastinal silhouette and hilar contours are normal . heart is normal in size . there is no pleural effusion or pneumothorax . no acute osseous abnormalities identified .	pa and lateral views of the chest provided . the lungs are clear . there is no focal consolidation , effusion , or pneumothorax . the cardiomediastinal silhouette is within normal limits . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .
the lungs are well expanded and clear . the cardiomediastinal silhouette , hilar contours and pleural surfaces appear normal . there is no pneumothorax or pleural effusion . the visualized bony structures are unremarkable .	frontal and lateral views of the chest were obtained . the lungs are well expanded and clear . there is no focal consolidation . no pleural effusion or pneumothorax . the heart is top normal in size with normal mediastinal and hilar contours . no radiopaque foreign body .	frontal and lateral radiographs of the chest demonstrate normal heart size and normal cardiomediastinal contours . the lungs are clear without consolidation . no pleural effusion or pneumothorax . no radiopaque foreign body .
the heart is normal in size . the hilar and mediastinal contours are within normal limits . the lungs are clear . there is no focal consolidation , pleural effusion or pneumothorax . visualized ossesous structures are grossly intact .	pa and lateral views of the chest . the lungs are clear without focal consolidation . no pleural effusion or pneumothorax . the cardiomediastinal silhouette is unremarkable . heart is normal in size . the hilar contours are within normal limits .	the lungs are well expanded and clear . no focal consolidation is seen . heart is normal in size . hilar contours are unremarkable . there is no pleural effusion or pneumothorax .
compared with the prior radiograph , there is a persistent left pleural effusion with over lying left basilar atelectasis and a newly identified right pleural effusion , evidenced by blunting of the costophrenic angles on the lateral view . there is no focal consolidation concerning for pneumonia or pneumothorax . unchanged median sternotomy wires , mediastinal clips , and right ij sheath .	the right ij central line tip is in the low svc . there is no pneumothorax . there are small bilateral pleural effusions , slightly increased from <unk> . there is atelectasis at the left base . there is no pneumothorax . the cardiac size is normal .	compared with the prior radiograph , there has been interval removal of the right ij central venous catheter . there are small bilateral pleural effusions with adjacent atelectasis . no pneumothorax , focal consolidation , or pulmonary edema . heart size is normal . cardiomediastinal silhouette is unchanged . median sternotomy wires are intact . surgical clips are noted in the left upper quadrant .
in comparison with the study of <unk> , there is little interval change . again there is substantial enlargement of the cardiac silhouette with some elevation of pulmonary venous pressure . opacification at the right base is consistent with volume loss in the lower lung and pleural effusion . mild atelectatic changes persist at the left base . monitoring and support devices remain in good position .	comparison is made to previous study from <unk> . there is an endotracheal tube whose distal tip is <num> cm above the carina , appropriately sited . there is a right ij central line with distal lead tip in the distal svc . there is a nasogastric tube whose tip and side port are below the ge junction . there is unchanged enlargement of cardiac silhouette . there is prominence of the pulmonary interstitial markings suggestive of pulmonary edema . there is a right basilar opacity which is likely due to atelectasis and pleural effusion . there is no pneumothorax . there are no pneumothoraces .	in comparison with the study of <unk> , the monitoring and support devices remain in place . continued enlargement of the cardiac silhouette with mild vascular congestion . opacification at the right base is consistent with pleural effusion and compressive atelectasis . the left base is essentially clear .
assessment is limited due to anteroposterior projection and positioning . allowing for these limitations : compared with the previous exam there appears to be worsening cardiomegaly , which now is moderate to severe allowing for limitations of this ap view . minimal interstitial edema is present . there is also a new right middle lobe opacity the , confirmed in the lateral view . no other focal opacities are identified . bilateral pleural effusions , left worse than right is present . there is no evidence of pneumothorax . left - sided two - lead pacemaker is reidentified , with one lead terminating in the right atrium and the other in the right ventricle . the aortic valve prosthesis appears unchanged . severe degenerative changes of the ac <unk>oints is again seen .	ap and lateral views of the chest provided . left chest wall pacer device is again seen with leads extending to the region of the right atrium and right ventricle . midline sternotomy wires are again noted . lung volumes are low . there is a small left pleural effusion with associated compressive atelectasis . there is mild interstitial pulmonary edema . no pneumothorax . cardiac silhouette is enlarged . no definite signs of congestion . degenerative changes at the right shoulder are noted . Missing: right middle lobe opacity, right pleural effusion	the lungs are well expanded . there is a small left pleural effusion with adjacent atelectasis . there is also a retrocardiac opacity which is likely atelectasis . there is also a small right pleural effusion . there is no pneumothorax . there is mild interstitial pulmonary edema . moderate cardiomegaly is unchanged . mediastinal and hilar contours are unremarkable . a left pectoral pacemaker is seen with leads in unchanged position in the right atrium and right ventricle . a left chest wall pacemaker is seen with leads in the expected positions of the right atrium and right ventricle .

Figure 4.4: Exemplar generated medical reports of our approaches (w/ and w/o Interpreter) for *normal* and *disease* cases. We highlighted the text that does not exist in the ground-truth report with pink and yellow colors. The highlighted pink-colored text indicates the topic exists in the chest X-ray images but does not mention in the ground-truth reports. In contrast, the highlighted yellow-colored text shows inaccurate information or cannot be confirmed from both chest X-ray images and ground-truth reports. For missing disease cases, we explicitly list them out in red-colored text.

Chapter 5

Conclusion and Outlook

Radiology report generation aims to detect abnormality in medical scans and write its findings with condensed sentences describing the diseases. It plays a crucial role in modern hospital systems and serves as the medium to communicate between doctors and patients. Overall, this research aimed to develop an end-to-end model that accurately diagnoses diseases and generates meaningful medical reports to assist physicians or radiologists. Based on the quantitative and qualitative analysis of the generated reports, it can be concluded that the proposed model can increase the quality of the generated reports compared to existing methods.

Particularly, in Chapter 2, we discussed the existing efforts in image captioning, report generation, and other related areas. From our observation, the existing efforts come with some advantages; however, they also have certain limitations, including difficulties in generating long documents, no parallelism, scalability issue, limited clinical accuracy. These limitations clearly put a barrier in bringing such models into production. Therefore, new medical report generation systems must address these challenges efficiently.

To address the parallelism and clinical accuracy issue, in Chapter 3, we presented an enriched disease embedding module that can polarize visual features into different directions called states. This allows the visual features to be semantic and serves as meaningful inputs for the report generation task. By transforming visual features into different states, we find that the EDDIE block can perform disease classification similar to a multi-label classifier with

comparable disease detection performance. Also, in this chapter, we propose to use the Transformer model to replace the existing hierarchical LSTM architectures for the medical report generation task. This comes with the benefit of parallelism and resolves long-range dependency issues in long documents. Our experimental results show that the Eddie-Transformer model reduces memory consumption and generates more fluent reports compared to RNN-based counterparts.

While Chapter 3 resolves the parallelism with an increase in clinical accuracy, it leaves some potential clinical inconsistencies between the detected diseases from medical images and the content of the generated reports. Moreover, as addressed in Chapter 2, our work only focuses on using a single input image, leaving other vital factors unchecked, such as the clinical history of patients. Therefore, in Chapter 4, we expand the previous chapter by proposing a complete model that works with different settings such as multi-view images or patients' clinical information. Furthermore, we propose a novel interpreter module that acts as a guidance network to fine-tune the generated reports to enhance the consistency of the classifier outputs and the generated reports. More importantly, the proposed interpreter is fully differentiable, eliminating the need for reinforcement learning when fine-tuning the reports. Our experiments, including human evaluation, show that the proposed approach can generate more accurate medical reports than other works by a large margin. This result emphasizes the possibility of designing an accurate medical report generation system.

With the promising results tested across different datasets from Chapter 3 and 4, we believe that the model can be applied to a wide range of related problems such as CT scans or brain imaging. We also believe that the proposed Interpreter module can help improve the quality of the generated text in other problems such as image captioning by enhancing factual information in generated captions. For future works, researchers may consider integrating disease locations into the framework to explicitly localize diseases better, further improving the quality of medical reports. Since our contextual information is limited to clinical indications provided by medical doctors, researchers can also

develop a time-series module to link previous radiology studies to the patient's current study to understand disease development better.

We hope that the methods presented in our work can inspire future methods in image captioning, medical report generation, and disease detection. We also hope that the proposed research can help future research take one step closer to real-world usage and assist physicians worldwide in their jobs.

References

[1] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[2] P. Barbaste, J. Lambert, J. Estorc, *et al.*, "Value of the lateral chest x-ray in the diagnosis of traumatic pleural effusions," in *Annales francaises d'anesthesie et de reanimation*, vol. 3, 1984, pp. 189–193.

[3] W. Boag, T.-M. H. Hsu, M. Mcdermott, G. Berner, E. Alesentzer, and P. Szolovits, "Baselines for chest x-ray report generation," in *NeurIPS Workshop on Machine Learning for Health*, 2020, pp. 126–140.

[4] P. J. Bossart, L. Brunsdale, M. Hughes, *et al.*, "The lateral chest x-ray: Is it necessary for emergency department patients?" *Emergency Radiology*, vol. 4, no. 1, pp. 26–29, 1997.

[5] C. Castillo, T. Steffens, L. Sim, and L. Caffery, "The effect of clinical information on radiology reporting: A systematic review," *Journal of Medical Radiation Sciences*, vol. 68, no. 1, pp. 60–74, 2021.

[6] E. M. Chamorro, A. D. Tascón, L. I. Sanz, S. O. Vélez, and S. B. Nacenta, "Radiologic diagnosis of patients with covid-19," *Radiologıéa (English Edition)*, vol. 63, no. 1, pp. 56–73, 2021.

[7] M. Chen, A. Radford, R. Child, *et al.*, "Generative pretraining from pixels," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1691–1703.

[8] X. Chen, H. Fang, T.-Y. Lin, *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[9] J. Cohen *et al.*, "Covid-19 image data collection," *ArXiv preprint*, vol. abs/ 2003.11597, 2020.

[10] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] C. for Disease Control, Prevention, *et al.*, "Strategies to mitigate healthcare personnel staffing shortages," 2020.

[13] J. Donahue, L. Anne Hendricks, S. Guadarrama, *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[14] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," in *Advances in neural information processing systems*, 1996, pp. 493–499.

[15] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125–4134.

[16] A. Gasimova, G. Seegoolam, L. Chen, P. Bentley, and D. Rueckert, "Spatial semantic-preserving latent space learning for accelerated dwi diagnostic report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 333–342.

[17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[18] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence?: Problems of tokenisation," 1994.

[19] S. Guendel *et al.*, "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *Iberoamerican Congress on Pattern Recognition*, Springer, 2018, pp. 757–765.

[20] M. P. Hartung, I. C. Bickle, F. Gaillard, and J. P. Kanne, "How to create a great radiology report," *RadioGraphics*, vol. 40, no. 6, pp. 1658–1670, 2020.

[21] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, *Gradient flow in recurrent nets: The difficulty of learning long-term dependencies*, 2001.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] J.-H. Huang, C.-H. H. Yang, F. Liu, *et al.*, "Deepopht: Medical report generation for retinal images via deep models and visual explanation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2442–2452.

[24] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[25] J. Irvin, P. Rajpurkar, M. Ko, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI Conference on Artificial Intelligence*, 2019.

[26] A. M. Ittyachen, A. Vijayan, and M. Isac, "The forgotten view: Chest x-ray-lateral view," *Respiratory medicine case reports*, vol. 22, pp. 257–259, 2017.

[27] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest x-ray reports," *arXiv preprint arXiv:2004.12274*, 2020.

[28] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[29] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, *et al.*, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[30] J. Kissane, J. A. Neutze, and H. Singh, "Lateral chest radiograph," in *Radiology Fundamentals*, Springer, 2020, pp. 63–68.

[31] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 317–325.

[32] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[33] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *arXiv preprint arXiv:1805.08298*, 2018.

[34] ——, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6666–6673.

[35] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3617–3625.

[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[37] T. Lin *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[38] G. Liu, T.-M. H. Hsu, M. McDermott, *et al.*, "Clinically accurate chest x-ray report generation," in *Machine Learning for Healthcare Conference*, PMLR, 2019, pp. 249–269.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[40] J. Lovelace and B. Mortazavi, "Learning to generate clinically coherent chest x-ray reports," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 1235–1243.

[41] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[42] Z. Lu, K. Deb, and V. N. Boddeti, "Muxconv: Information multiplexing in convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 044–12 053.

[43] Z. Lu, I. Whalen, Y. Dhebar, *et al.*, "Multi-objective evolutionary design of deep convolutional neural networks for image classification," *IEEE Transactions on Evolutionary Computation*, 2020.

[44] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[45] R. McLoughlin, C. So, R. Gray, and R. Brandt, "Radiology reports: How much descriptive detail is enough?" *AJR. American journal of roentgenology*, vol. 165, no. 4, pp. 803–806, 1995.

[46] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 746–751.

[47] T. Nishino, R. Ozaki, Y. Momoki, *et al.*, "Reinforcement learning with imbalanced dataset for data-to-text medical report generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2223–2236.

[48] P. Obara, M. Sevenster, A. Travis, Y. Qian, C. Westin, and P. J. Chang, "Evaluating the referring physician's clinical history and indication as a means for communicating chronic conditions that are pertinent at the point of radiologic interpretation," *Journal of digital imaging*, vol. 28, no. 3, pp. 272–282, 2015.

[49] F. Osman and I. Williams, "Should the lateral chest radiograph be routinely performed?" *Radiography*, vol. 20, no. 2, pp. 162–166, 2014.

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[51] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8347–8356.

[52] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[53] P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[54] E. Ranjan, S. Paul, S. Kapoor, A. Kar, R. Sethuraman, and D. Sheet, "Jointly learning convolutional representations to compress radiological images and classify thoracic diseases in the compressed domain," in *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, 2018, pp. 1–8.

[55] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[56] I. Satia, S. Bashagha, A. Bibi, R. Ahmed, S. Mellor, and F. Zaman, "Assessing the accuracy and certainty in interpreting chest x-rays in the medical division," *Clinical medicine*, vol. 13, no. 4, p. 349, 2013.

[57] Y. Shibata, T. Kida, S. Fukamachi, *et al.*, "Byte pair encoding: A text compression scheme that accelerates pattern matching," 1999.

[58] S. Singh, S. Karimi, K. Ho-Shon, and L. Hamey, "From chest x-rays to radiology reports: A multimodal machine learning approach," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2019, pp. 1–8.

[59] P. Srinivasan, D. Thapar, A. Bhavsar, and A. Nigam, "Hierarchical x-ray report generation via pathology tags and multi head attention," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[60] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[61] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 035–13 045.

[62] M. Uimonen, I. Kuitunen, J. Paloneva, A. P. Launonen, V. Ponkilainen, and V. M. Mattila, "The impact of the covid-19 pandemic on waiting times for elective surgery patients: A multicenter study," *Plos one*, vol. 16, no. 7, e0253875, 2021.

[63] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[64] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[65] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[66] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[67] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[68] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "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*, 2018, pp. 9049–9058.

[69] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[70] J. R. Wilcox, "The written radiology report," *Applied Radiology*, vol. 35, no. 7, p. 33, 2006.

[71] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 033–10 041.

[72] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2019, pp. 673–680.

[73] K. Xu, J. Ba, R. Kiros, *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, PMLR, 2015, pp. 2048–2057.

[74] Y. Xue, T. Xu, L. R. Long, *et al.*, "Multimodal recurrent model with attention for automated radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 457–466.

[75] C. Yin *et al.*, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *ICDM*, 2019, pp. 728–737.

[76] C. Yin, B. Qian, J. Wei, *et al.*, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *2019 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2019, pp. 728–737.

[77] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[78] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 721–729.

[79] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 8–15.

[80] H. Zhang *et al.*, "Resnest: Split-attention networks," *arXiv*, 2020.

[81] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 910–12 917.

[82] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

Appendix A

Supplementary Material

A.1 Implementation Details

A.1.1 Dataset Preprocess and Splits

As a preprocessing step, all tokens in the medical reports are converted to lowercase. We split both MIMIC-CXR and Open-I datasets with a standard ratio of 70:10:20% for training, validation, and testing purposes, respectively.

For the MIMIC-CXR dataset, in addition to the 14 disease labels extracted from the CheXpert labeler [25], we also obtain the top-100 high-frequency noun-phrases of the dataset as our additional disease-related topics using Spacy[1]. In total, a list of 114 disease-related topics are acquired as the binary targets of the induced classification task. Each disease label is either *positive* (including *uncertain* or *exist*) or *negative* (including *unmentioned* or *non-exist*). We also ensure no patient overlap across the train and test sets to avoid data leakage.

For the Open-I dataset, since this dataset does not have any ground-truth disease labels, the 14 diseases extracted from the MIMIC-CXR dataset is used here, together with the top-100 high-frequency noun-phrases obtained in the Open-I dataset. This again forms the 114 disease-related topics as binary classification targets.

[1]Spacy is an NLP industrial open-source project: https://spacy.io

A.1.2 Optimizer

Adam with decoupled Weight decay regularization - AdamW [39] is used throughout our experiments. Specifically, the initial learning rates are set to $\eta = 3 \times 10^{-4}$ and $\beta = (0.9, 0.999)$, respectively; $L2$ regularization weight decay is set to $\lambda = 10^{-2}$. Models are trained for 50 epochs, with a learning rate schedule $\eta = 3 \times 10^{-5}$ after the 25-th epoch.

A.1.3 Data Input and Augmentation

Input images are rescaled to 256×256 pixels; the following image transformations are randomly applied to account for overfitting: image rotating, color jittering, and horizontal flipping. At the end of the image encoder, we set a dropout rate to 0.1. For the text inputs, we construct a vocabulary containing the top 900 high-frequency words and 100 byte-pair-encoding tokens[2] or BPE, to avoid the out-of-vocabulary scenario. It gives rise to a vocabulary of 1,000 words and tokens covering approximately 99% of the total words and tokens in the datasets.

A.1.4 Text Encoder and Interpreter

The transformer encoder of [63] is used as our text encoder, which consists of a multi-head self-attention layer and a feed-forward layer. We set the number of heads NumHeads = 8, and the number of neurons in the feed-forward layer FwdDim = 256. The embedding dimension is set to $e = 256$. Finally, a dropout rate of 0.1 is applied here to avoid overfitting.

A.1.5 Generator

The configuration of the generation module is slightly different from the text encoder module. It consists of 12 masked multi-head self-attention layers and feed-forward layers, which allows a large receptive field in generating medical reports. The number of heads is set to NumHeads = 1. The number of neurons

[2]SentencePiece is a vocabulary builder open-source project developed by Google to facilitate BPE tokenization, at https://github.com/google/sentencepiece

in the feed-forward layer is FwdDim = 256. The embedding dimension is set to $e = 256$. Similar to the text encoder, a dropout rate of 0.1 is applied to avoid overfitting.

A.1.6 Classifier

It is observed that both chest X-ray datasets are highly imbalanced w.r.t. the disease-related topics: positive cases are often significantly lower than the negative cases [40]. Therefore, the classification network often has a very low confidence score p and favors a negative prediction. It may increase the false negative cases that could be very costly in the medical context. To account for this issue, a threshold is adopted to the Eq. 5 (in the main text), as

$$p = \begin{cases} 1, & p > \text{threshold} \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

The threshold value is determined using grid-search on the validation dataset. The search range is from 0.1 to 0.5. The grid size is 0.05. In the full version (MV+T+I) of our approach, it is set to threshold = 0.25 for the MIMIC-CXR dataset and threshold = 0.15 for the Open-I dataset. See Table A.1 and A.2 for more detail.

A.2 The top noun-phrases

This section lists the top-100 high-frequency noun phrases as our additional disease-related topics for both the MIMIC-CXR and the Open-I dataset. Since the noun-phrases are automatically extracted using Spacy, some noun-phrases may overlap and repeat several times. The top noun phrases separated by semicolons are listed below.

A.2.1 Top-100 noun-phrases of MIMIC-CXR dataset

pneumothorax; the lungs; no pleural effusion; the chest; no pneumothorax; pleural effusion; no focal consolidation; the cardiomediastinal silhouette; normal limits; ¡unk; heart size; atelectasis; lungs; the cardiac silhouette; pa; no evidence; pneumonia; focal consolidation; the heart; effusion; lateral views;

Method	Threshold	0.05	0.1	0.15	0.2	0.25	0.3
Ours (SV)	BLEU-1	0.355	**0.470**	0.449	0.427	0.389	0.370
	BLEU-2	0.229	**0.320**	0.315	0.294	0.266	0.251
	BLEU-3	0.156	**0.227**	0.226	0.209	0.187	0.177
	BLEU-4	0.107	0.162	**0.164**	0.150	0.133	0.127
	METEOR	**0.200**	0.190	0.187	0.181	0.174	0.169
	ROUGE-L	0.343	0.382	**0.395**	0.381	0.384	0.378
Ours (MV)	BLEU-1	0.344	0.433	**0.476**	0.464	0.433	0.405
	BLEU-2	0.225	0.297	**0.329**	0.321	0.296	0.273
	BLEU-3	0.152	0.213	**0.235**	0.230	0.210	0.194
	BLEU-4	0.104	0.156	**0.171**	0.168	0.152	0.142
	METEOR	**0.204**	0.199	0.198	0.193	0.182	0.176
	ROUGE-L	0.338	0.377	0.386	**0.391**	0.382	0.374
Ours (MV+T)	BLEU-1	0.364	0.449	**0.500**	0.498	0.481	0.436
	BLEU-2	0.251	0.323	0.371	**0.375**	0.363	0.331
	BLEU-3	0.181	0.245	0.292	**0.299**	0.290	0.267
	BLEU-4	0.133	0.190	0.236	**0.245**	0.239	0.222
	METEOR	0.209	0.212	**0.213**	**0.213**	0.209	0.199
	ROUGE-L	0.361	0.408	0.437	0.446	**0.451**	0.448
Ours (MV+T+I)	BLEU-1	0.401	0.473	**0.523**	0.507	0.484	0.464
	BLEU-2	0.280	0.346	**0.393**	0.384	0.370	0.353
	BLEU-3	0.207	0.268	**0.313**	0.311	0.301	0.286
	BLEU-4	0.157	0.215	0.257	**0.260**	0.252	0.241
	METEOR	0.222	0.226	**0.227**	0.221	0.217	0.212
	ROUGE-L	0.390	0.432	0.453	0.454	**0.458**	0.453

Table A.1: Optimal threshold search on the Open-I validation dataset. Bounded columns indicate our choice for the optimal threshold.

Method	Threshold	0.1	0.15	0.2	0.25	0.3	0.35
Ours (SV)	BLEU-1	0.310	0.363	0.408	**0.445**	0.407	0.353
	BLEU-2	0.208	0.242	0.267	**0.289**	0.260	0.226
	BLEU-3	0.145	0.170	0.187	**0.200**	0.179	0.156
	BLEU-4	0.105	0.123	0.136	**0.144**	0.128	0.112
	METEOR	**0.217**	0.212	0.199	0.186	0.170	0.156
	ROUGE-L	0.297	0.312	**0.319**	0.318	0.312	0.309
Ours (MV)	BLEU-1	0.316	0.366	0.416	**0.451**	0.406	0.360
	BLEU-2	0.213	0.245	0.274	**0.292**	0.261	0.229
	BLEU-3	0.150	0.172	0.192	**0.201**	0.178	0.157
	BLEU-4	0.109	0.125	0.139	**0.144**	0.127	0.112
	METEOR	**0.218**	0.211	0.200	0.187	0.171	0.156
	ROUGE-L	0.301	0.314	**0.321**	**0.321**	0.314	0.307
Ours (MV+T)	BLEU-1	0.370	0.418	0.462	0.495	**0.496**	0.465
	BLEU-2	0.268	0.304	0.337	**0.362**	0.361	0.338
	BLEU-3	0.203	0.232	0.260	0.280	**0.281**	0.263
	BLEU-4	0.159	0.184	0.207	0.225	**0.226**	0.213
	METEOR	**0.243**	0.240	0.234	0.224	0.216	0.206
	ROUGE-L	0.355	0.376	0.385	0.390	**0.391**	0.388
Ours (MV+T+I)	BLEU-1	0.371	0.423	0.466	**0.497**	0.494	0.463
	BLEU-2	0.267	0.307	0.340	**0.364**	0.362	0.338
	BLEU-3	0.202	0.235	0.262	**0.282**	0.281	0.264
	BLEU-4	0.157	0.186	0.210	**0.227**	**0.227**	0.214
	METEOR	**0.240**	0.238	0.233	0.224	0.216	0.206
	ROUGE-L	0.356	0.376	0.389	**0.391**	**0.391**	0.390

Table A.2: Optimal threshold search on the MIMIC validation sub-dataset (1000 random validation samples). Bounded columns indicate our choice for the optimal threshold.

comparison; no acute osseous abnormalities; pulmonary edema; size; cardiomediastinal silhouette; the right hemidiaphragm; lung volumes; the mediastinal and hilar contours; the heart size; the previous radiograph; the patient; the carina; no free air; low lung volumes; consolidation; the pulmonary vasculature; frontal and lateral views; the cardiac and mediastinal silhouettes; the stomach; the prior study; bony structures; the aorta; mediastinal and hilar contours; no pulmonary edema; no pleural effusions; evidence; the right; the tip; the study; the right atrium; the left; the right lung; the thoracic spine; moderate cardiomegaly; pulmonary vasculature; the left lung; position; edema; the right lung base; the cardiomediastinal and hilar contours; no acute osseous abnormality; cardiac silhouette; the lung bases; patient; the left lung base; mediastinal contours; place; the lung volumes; imaged osseous structures; cardiomediastinal contours; pulmonary vascular congestion; mild cardiomegaly; unchanged position; appearance; the level; the lateral view; ¿; hilar contours; small bilateral pleural effusions; no large pleural effusion; normal size; infection; the thoracic aorta; vascular congestion; it; mild pulmonary edema; the diaphragm; heart; bibasilar atelectasis; tip; aspiration; the cardiac, mediastinal and hilar contours; no relevant change; the left lower lobe; the left hemidiaphragm; the mediastinal contours; the mid svc; mediastinal contour; study.

A.2.2 Top-100 noun-phrases of Open-I dataset

pneumothorax; normal limits; the lungs; pleural effusion; no pneumothorax; heart size; the heart; lungs; size; no pleural effusion; the cardiomediastinal silhouette; xxxx; mediastinum; no focal consolidation; pulmonary vascularity; contour; focal airspace disease; the thoracic spine; the heart size; heart; large pleural effusion; pulmonary vasculature; mediastinal contours; the mediastinum; focal consolidation; the spine; cardiomediastinal silhouette; degenerative changes; no pleural effusions; normal heart size; effusion; pneumothoraces; no focal airspace consolidation; visualized osseous structures; evidence; pulmonary xxxx; xxxx xxxx; the xxxx; low lung volumes; no evidence; consolidation; both lungs; mediastinal contour; bony structures; the chest; appearance;

effusions; osseous structures; masses; normal size; no acute bony abnormality; the thorax; no focal areas; acute abnormality; the interval; specifically, no evidence; the aorta; mild degenerative changes; the cardiac silhouette; pulmonary edema; the thoracic aorta; atelectasis; the skeletal structures; cardiac and mediastinal contours; cardio mediastinal silhouette; soft tissues; no large pleural effusion; infiltrate; no definite pleural effusion; the trachea; no focal airspace disease; no acute bony findings; pleural spaces; no focal air space opacity; no focal alveolar consolidation; edema; adenopathy; no acute bony abnormalities; no effusion; the cardiomediastinal contours; lung volumes; no typical findings; no visible pneumothorax; no pneumonia; the diaphragm; nodules; a pneumonia; suspicious pulmonary opacity; no focal infiltrate; clear lungs; thoracic spondylosis; configuration; no acute abnormality; surgical clips; upper limits; the xxxx examination; bony thorax; stable cardiomediastinal silhouette; bronchovascular crowding; the left lung base;

A.3 Discussion on the Classifier module

From Table A.3, we can observe that the classifier performance improves with additional input data (SV,MV,MV+T). With the Interpreter module, the recall score increases, which is encouraging in the medical context as being discussed in the main text. Additionally, it is clear from Table A.3 that the performance of the generator relies on the performance of the classifier module. In other words, the generator can only generate high-quality medical reports when the classifier can detect diseases accurately.

Note that the scores showed in Table A.3 is the commonly used image classification metrics. In contrast, the scores shown in the main paper are obtained by reading the generated text, classifying it into disease labels (multi-label text classifier) via the CheXpert labeler, then comparing them with the ground-truth disease labels. Therefore, the scores of the classifier (measuring the ability to classify diseases from images) shown in Table A.3 and the scores of the generator (measuring the quality of the generated medical reports) shown in the main paper are different.

Dataset	Methods	Accuracy	AUC	F1	Precision	Recall
Open-I	Ours (SV)	0.929	0.740	0.176	0.146	0.232
	Ours (MV)	0.932	**0.795**	0.214	0.253	0.252
	Ours (MV+T)	**0.934**	0.777	0.217	**0.284**	0.233
	Ours (MV+T+ I)	0.925	0.784	**0.236**	0.236	**0.290**
MIMIC	Ours (SV)	0.873	0.822	0.389	0.392	0.451
	Ours (MV)	0.875	0.829	0.399	0.397	0.465
	Ours (MV+T)	**0.890**	**0.865**	0.464	**0.458**	0.502
	Ours (MV+T+ I)	0.880	0.863	**0.475**	0.431	**0.551**

Table A.3: This table shows the classifier module's performance of our proposed model on the 14 common diseases. The improvement of our classifier's module is consistent with the reported performance from the automated metrics (fluency and accuracy) obtained from the generated reports.