

Automatic Generation of Radiological Medical Image Report

ZHANYU WANG

SID: 490450411

Supervisor: Dr. Luping Zhou

Associate Supervisor: Dr. Wanli Ouyang

A Thesis Submitted in Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

School of Electrical and Information Engineering
Faculty of Engineering and Information Technology
The University of Sydney

23 April 2024

To my family.

Statement of Originality

This thesis is submitted to the School of Electrical and Information Engineering, Faculty of Engineering and Information Technology, the University of Sydney in fulfilment of the requirements for the degree of Doctor of Philosophy. I certify that the content of this thesis is my own work and this thesis has not been submitted for any degree or other purposes. I also certify that all the assistance received in preparing this thesis and sources have been acknowledged.

Zhanyu Wang

Date

Authorship Attribution Statement

This thesis was conducted at the University of Sydney, under the supervision of Prof. Luping Zhou, between 2020 and 2023. The main results presented in this dissertation were first introduced in the following publications:

(1) **Z. Wang**, L. Zhou, L. Wang, X. Li. "A self-boosting framework for automated radiographic report generation" In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021: 2433-2442. Presented in Chapter 2. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

(2) **Z. Wang**, L. Zhou, L. Wang, X. Li. "Diagnostic Captioning by Cooperative Task Interactions and Sample-graph Consistency" *Under review*. Presented in Chapter 3. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

(3) **Z. Wang**, M. Tang, L. Wang, X. Li, L. Zhou. "A medical semantic-assisted transformer for radiographic report generation" In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2022: 655-664. Presented in Chapter 4. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

(4) **Z. Wang**, H. Han, L. Wang, X. Li, L. Zhou. "Automated radiographic report generation purely on transformer: A multicriteria supervised approach".

In *IEEE Transactions on Medical Imaging (T-MI)*. 2022, 41(10): 2803-2813. Presented in Chapter 5. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

(5) **Z. Wang**, L. Liu, L. Wang, L. Zhou. "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens" In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023: 11558-11567. Presented in Chapter 6. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

(6) **Z. Wang**, L. Liu, L. Wang, L. Zhou. "R2GenGPT: Radiology Report Generation with Frozen LLMs". In *Meta-Radiology*. 2023: 100033. Presented in Chapter 7. I designed the study, implemented the system, conducted the experiments and ablations, and wrote the draft manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Zhanyu Wang

Date

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Luping Zhou

Date

Acknowledgement

I extend my sincere gratitude to all individuals and organizations whose contributions were instrumental in the successful completion of my doctoral thesis.

First and foremost, I would like to express my heartfelt appreciation to my supervisor, Prof. Luping Zhou. Throughout my PhD journey, Prof. Zhou's unwavering support and guidance have been invaluable. Her patience, encouragement, and insightful suggestions played a pivotal role in helping me overcome research challenges and enhancing my skills. I am profoundly grateful for the trust and autonomy she afforded me, allowing me to delve into research topics I am passionate about. Prof. Zhou's dedication to academic excellence and her wealth of expertise have left an enduring impact on my academic and professional development.

I also want to extend my gratitude to my auxiliary supervisors, Prof. Lei Wang from the University of Wollongong and Prof. Lingqiao Liu from the University of Adelaide, for their generous support and guidance. Working under the tutelage of Prof. Wang and Prof. Liu was a privilege, and their mentorship provided me with invaluable insights into the field of machine learning.

I am thankful for the camaraderie and support of my friends, sorted alphabetically by last name: Dr. Zhenghao Chen, Mr. Tong Chen, Mr. Yiqun Duan, Ms. Erjian Guo, Ms. Yunyi Liu, Dr. Yinyan Liu, Ms. Xinhui Liu, Ms. Xinyu Liang, Mr. Rui Luo, Mr. Cheng Wen, Dr. Yiming Wu, Mr. Ling Yang, Mr. Xianghui Yang, Ms. Haoning Xi, and Mr. Peiqin Zhuang. The moments we shared during my time in Sydney were cherished, and your support has meant the world to me.

I want to acknowledge the unwavering support of my family, including my parents, my elder sister, her husband, and their entire family, who have been my pillars of strength throughout my academic journey. Additionally, I am deeply grateful for the love and support of my girlfriend, Shasha Peng, who has been a constant source of inspiration and encouragement.

In closing, I extend my heartfelt thanks to all who contributed to my doctoral research. Your support and guidance have been invaluable in shaping my academic and personal growth. I am truly appreciative of your pivotal roles in this important chapter of my life.

Abstract

Automated Radiographic Report Generation (R2Gen) is a critical task aimed at the automatic generation of a free-text description for medical images, such as chest X-rays. This process entails producing a comprehensive and contextually relevant summary of the image’s content and clinical significance. Given its significant potential to alleviate the substantial workload faced by radiologists, R2Gen has garnered increasing attention from researchers worldwide. While significant progress has been made, R2Gen still faces several challenges: The first challenge lies in the generation of long texts. Many current methods employ an encoder-decoder architecture similar to image captioning for natural images. However, in contrast to a brief description in image captioning, R2Gen needs to generate lengthy paragraphs with intricate details. The second challenge lies in the fine-grained image distinctions. Radiological images, such as chest X-rays, often exhibit high similarity, making it challenging for the model to discern differentiating features between them. The third challenge lies in the bias in visual and textual data. Due to an over-representation of normal samples in the training data, the model’s learning process was biased towards these samples, limiting its ability to effectively detect abnormalities and anomalies within the dataset.

To tackle these challenges, in this thesis, we propose several innovative approaches to improve the quality and accuracy of radiographic report generation. First, We propose a self-boosting framework that synchronizes image and report features by simultaneously training an auxiliary image-text matching task alongside the main report generation task. This mutual training process allows

each task to refine and enhance the other, leading to more accurate report generation aligned closely with ground-truth reports. We also explore sample relationships between image and report features by introducing a sample-graph consistency loss as an additional regularization into the self-boosting framework. Subsequently, with the rise in popularity of the Transformer architecture, we devised a novel mechanism to embed the image-text matching task within a Transformer’s encoder-decoder structure, employing a pure Transformer network structure in the R2Gen task for the first time. Moreover, we investigate ways to incorporate external information to assist in the generation of reports. To this end, we have designed a Medical Concepts Generation Network to predict fine-grained semantic concepts and integrate them into the report generation process as guidance. Next, treating the current method as a “single expert”, we introduced a new diagnostic captioning framework, METransformer, to emulate the clinical scenario of “collaborative diagnosis by multiple experts”. The key design of METransformer is the introduction of learnable expert tokens and encouraging them to learn complementary representations using both linear and non-linear attentions. In the last, we propose a novel LLMs-based Radiology report generation framework, dubbed R2GenGPT, which is the first instance of harnessing pre-trained large language models (LLMs) for the R2Gen task. Experimental results demonstrate the effectiveness of our methods on two public datasets, showing its superior performance over multiple state-of-the-art image captioning and medical report generation methods.

Publication List

- (1) **Z. Wang**, L. Zhou, L. Wang, X. Li. "A self-boosting framework for automated radiographic report generation" In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021: 2433-2442. [**Chapter 2**]
- (2) **Z. Wang**, L. Zhou, L. Wang, X. Li. "Diagnostic Captioning by Cooperative Task Interactions and Sample-graph Consistency". *Under review*. [**Chapter 3**]
- (3) **Z. Wang**, M. Tang, L. Wang, X. Li, L. Zhou. "A medical semantic-assisted transformer for radiographic report generation". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2022: 655-664. [**Chapter 4**]
- (4) **Z. Wang**, H. Han, L. Wang, X. Li, L. Zhou. "Automated radiographic report generation purely on transformer: A multicriteria supervised approach". In *IEEE Transactions on Medical Imaging (T-MI)*. 2022, 41(10): 2803-2813. [**Chapter 5**]
- (5) **Z. Wang**, L. Liu, L. Wang, L. Zhou. "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023: 11558-11567. [**Chapter 6**]

- (6) **Z. Wang**, L. Liu, L. Wang, L. Zhou. "R2GenGPT: Radiology Report Generation with Frozen LLMs". In *Meta-Radiology*. 2023: 100033. **[Chapter 7]**

- (7) Y Liu, **Z. Wang**, D. Xu, L. Zhou. "Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder". In *International Conference on Information Processing in Medical Imaging (IPMI)*. 2023: 445-456.

- (8) M. Tang, **Z. Wang**, Z. Zeng, X. Li, L. Zhou. "Stay in Grid: Improving Video Captioning via Fully Grid-level Representation". In *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*. 2022: 3319-3332.

CONTENTS

Statement of Originality	v
Authorship Attribution Statement	vi
Acknowledgement	viii
Abstract	x
Publication List	xii
List of Figures	xviii
List of Tables	xxiv
Chapter 1 Introduction	1
1.1 Background	1
1.2 Outline	5
1.3 Contributions	7
Chapter 2 Literature Review	9
2.1 Image Captioning	9
2.2 Radiology report generation	11
2.3 Transformers	13
2.4 Image-Text Matching	14
2.5 Large language Models	15
Chapter 3 A Self-boosting Framework for Automated Radiographic Report Generation	16
3.1 Introduction	17
3.2 Method	19

3.2.1	Overview	19
3.2.2	Report Generation (RG) Branch	21
3.2.3	Image-Text Matching (ITM) Branch	24
3.2.4	Self-boosted Training	27
3.2.5	Remarks	28
3.3	Experiments	29
3.3.1	Datasets	29
3.3.2	Experimental Settings	30
3.3.3	Results and Discussion	30
3.4	Conclusions	34
Chapter 4 Diagnostic Captioning by Cooperative Task Interactions and Sample-graph Consistency		36
4.1	Introduction	37
4.2	Method	39
4.2.1	Overview	40
4.2.2	Report Generation (RG) Branch	42
4.2.3	Image-Text Matching (ITM) Branch	45
4.2.4	Self-boosted Training	48
4.2.5	Sample-graph Consistency Loss	49
4.3	Experiments	51
4.3.1	Datasets	51
4.3.2	Experimental Settings	52
4.3.3	Results and Discussion	53
4.4	Conclusions	64
Chapter 5 A Medical Semantic-Assisted Transformer for Radiographic Report Generation		65
5.1	Introduction	66
5.2	Method	67

5.2.1	Memory-augmented Sparse Attention of High-order Interaction	68
5.2.2	Image Encoder	70
5.2.3	Medical Concepts Generation Network	71
5.2.4	Report Decoder	72
5.3	Experiments	73
5.3.1	Data Collection	73
5.3.2	Experimental Settings	73
5.3.3	Results and Discussion	74
5.4	Conclusions	76
 Chapter 6 Automated Radiographic Report Generation Purely On Transformer: A Multi-criteria Supervised Approach 77		
6.1	Introduction	78
6.2	Method	81
6.2.1	Structure of the Proposed Model	82
6.2.2	Multi-criteria Objectives	85
6.3	Experiments	89
6.3.1	Datasets	89
6.3.2	Experimental Settings	91
6.3.3	Results and Discussion	92
6.4	Conclusion	98
 Chapter 7 METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens 101		
7.1	Introduction	102
7.2	Method	104
7.2.1	Multi-expert ViT Encoder	105
7.2.2	Multi-expert Bilinear Attention Encoder	107
7.2.3	Multi-expert Bilinear Attention Decoder	108
7.2.4	Objective Function	109

7.2.5	Expert Voting strategy	110
7.3	Experiments	111
7.3.1	Datasets	111
7.3.2	Experimental Settings	111
7.3.3	Main Results	113
7.3.4	Ablation Study	115
7.3.5	Qualitative analysis.	117
7.4	Conclusions	119
Chapter 8	R2GenGPT: Radiology Report Generation with Frozen LLMs	120
8.1	Introduction	120
8.2	Methodology	123
8.3	Experiments	126
8.3.1	Data Collection	127
8.3.2	Experimental Settings	127
8.3.3	Results and Discussion	129
8.4	Conclusions	131
Chapter 9	Conclusions	133
9.1	Future Outlook	135
References		138

List of Figures

- 1.1 An example of Chest X-ray data used in this thesis. 3
- 1.2 An example of a Chest X-ray alongside a corresponding radiological report in Chinese. 4
- 3.1 An overview of the proposed framework comprising of a medical report generation branch (RG-branch, the gray part) and an auxiliary image-text matching branch (ITM-branch, the green part). RG-branch consists of two components: a region-level image encoder (shared with ITM-branch) to extract visual features from medical images, and a hierarchical LSTM to generate diagnose reports. ITM-branch consists of an image encoder (shared with RG-branch) and a report encoder based on Sentence-BERT. It compares the similarity of visual and text features for match. The two branches are designed to have cooperative interactions during training. ITM-branch provides RG-branch text-correlated visual features for report generation and its report encoder for measuring the feature level loss of the generated reports. In turn, with the improved generated reports, RG-branch supplies ITM-branch harder and harder samples to force the latter to evolve for better feature learning. In this way, the whole model is self-boosted. 20
- 3.2 Examples of the generated report on IU-xray(upper part) and COV-CTR dataset (lower part). We compare our results with the ground truth report and the other two methods: Show-Tell (Vinyals *et al.*, 2015) and Grounded (Zhou *et al.*, 2020). Highlighted sentences are the key

- information in the ground truth report that is only generated by our method. 32
- 4.1 Compared with the common encoder-decoder structure used for report generation (a), our self-boosting framework for diagnostic captioning (b) jointly trains two tasks: the main task of report generation (RG), and the auxiliary task of image-text matching (ITM), built as two branches of a deep learning model. The RG branch (indicated by the solid line) takes an image as the input and outputs the generated report. The ITM branch (indicated by the dash line) takes an image-report pair as the input and predicts whether they are a true match. The RG-branch and the ITM branch share the same image encoder. They gradually help improve each other through their cooperative interactions. 38
- 4.2 An overview of the network structure of the proposed framework with a Transformer-based decoder. It comprises of a medical report generation branch (RG-branch, the upper part) and an auxiliary image-text matching branch (ITM-branch, the lower part). RG-branch consists of two components: a region-level image encoder (shared with ITM-branch) to extract visual features from medical images, and a memory-augmented Transformer decoder to generate diagnose reports. ITM-branch consists of an image encoder (shared with RG-branch) and a report encoder based on Sentence-BERT. It compares the similarity of visual and text features for match. The two branches are designed to have cooperative interactions during training. ITM-branch provides RG-branch text-correlated visual features for report generation and its report encoder for measuring the feature level loss of the generated reports. In turn, with the improved generated reports, RG-branch supplies ITM-branch harder and harder samples to force the latter to evolve for better feature learning. In this way, the whole model is self boosted. 41

- 4.3 Illustration of our sample-graph consistency loss. Blue and green indicate the image modality and the report modality, respectively. The images and the reports are mapped into their respective feature spaces by the image encoder h_I and the report encoder h_T . In the feature spaces, both the sample-graph $\mathbf{W}_I$ of the images and the sample-graph $\mathbf{W}_T$ of the reports are constructed. The graph nodes sharing the same symbol in $\mathbf{W}_I$ and $\mathbf{W}_T$ belong to one image-report pair and the edge length between two nodes (i.e., samples) of a graph represents their distance in the embedding space. Initially, in the image feature space, the embeddings of two Chest-Xray images may be very close due to the fine-grained visual difference, despite the essential difference in their corresponding reports. Our sample-graph consistency loss takes the sample-graph $\mathbf{W}_T$ of the ground-truth reports as the supervision, and minimizes the difference between the two sample-graphs $\mathbf{W}_T$ and $\mathbf{W}_I$ so that two similar images with different ground-truth reports could stay apart in the feature embedding space. 50
- 4.4 Examples of the generated reports on IU-xray dataset. We compare our results with the ground truth report, our baseline using RG-branch only with the cross-entropy loss (denoted as RG-branch(CE only)), and another three competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports represent incorrectly predicted sentences, sentences not exactly consistent with ground-truth, and sentences that completely match the ground-truth, respectively. 56
- 4.5 Examples of the generated reports on MIMIC-CXR dataset. We compare our results with the ground truth report, our baseline using RG-branch only with the cross-entropy loss (denoted as RG-branch(CE only)), and another three competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*,

- 2020b). The highlighted sentences with red, blue and green color in the generated reports represent incorrectly predicted sentences, sentences not exactly consistent with ground-truth, and sentences that completely match the ground-truth, respectively. 57
- 5.1 An overview of the proposed framework, which comprises an Image Encoder, a Medical Concepts Generation Network (MCGN) and a Report Decoder. The transformer encoder and decoder are embedded with a Memory-augmented Sparse nonlinear Attention (MSA) to capture the higher-order interactions between the input fine-grained image features. Meanwhile, we inject a medical concepts generator to provide semantic information to facilitate report generation. 68
- 5.2 An example report generated by the proposed model. The correctly predicted medical concepts and the key information in the report are marked red. 75
- 6.1 An overview of the proposed framework which comprises of a vision transformer encoder and a memory augmented text transformer decoder. The whole model is supervised by three criteria: 1) a term weighting loss for the main task of report generation (RG loss), 2) an image-text matching loss with temporal-weighting (ITM loss), which facilitates report generation by enforcing the model to learn strongly correlated visual and textual features, and 3) a multi-label classification loss (MLC loss) that regularises the report generation by ensuring the model to make correct diagnostic prediction. 81
- 6.2 (a) The shift-right mask: blue indicates “1” and white indicates “0”. (b) x -axis is the masking step, while y -axis is the ratio of text information (i.e., S_{tr}) in the similarity score. Due to the shift-right masked self-attention, this ratio increases with steps. Accordingly, we assign higher weight (green curve) to the similarity score of later steps to compute image-text matching loss. 86

- 6.3 Examples of the generated reports on IU-xray dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth. 97
- 6.4 Examples of the generated reports on MIMIC-CXR dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth. 98
- 6.5 The failure cases for IU-xray dataset (upper case) and MIMIC-CXR dataset (lower case). Sentences marked in red indicate frequently occurring sentences in the report 99
- 7.1 An overview of our proposed METransformer, which includes an Expert Transformer Encoder and an Expert Transformer Decoder. The expert transformer encoder captures fine-grained visual patterns by exploring 1st and 2nd order interactions between input visual tokens and M expert tokens. The expert transformer decoder produces M diverse candidate reports guided by each expert token through an Adjust block. Noted that the “Embed” layer includes a linear projection followed by a ReLU activation function. 105
- 7.2 Bleu_4 and CIDEr scores by using different numbers of expert tokens on IU-Xray and MIMIC-CXR dataset. 116

7.3 Attention visualization of expert tokens on the image.	118
7.4 An example of the generated reports and their attention-mapping visualization of three key medical terms from BASELINE and ours METransformer. For better illustration, different colors in the generated reports highlight different medical information.	119
8.1 An overview of our proposed R2GenGPT. The input tokens for the Large Language Model (LLM) are sequentially concatenated, consisting of visual tokens, prompt tokens, and report tokens. A token mask of -100 indicates that those particular tokens are excluded from auto-regressive training, while a mask of 1 signifies inclusion in auto-regressive training.	124
8.2 Three proposed alignment methods. (a) Shallow Alignment: Training only the Linear Layer. (b) Deep Alignment: Training both the Linear Layer and all parameters of the Visual Encoder. (c) Delta Alignment: Training the Linear Layer and a small subset of incremental parameters of the Visual Encoder.	125
8.3 Examples of the generated report on MIMIC-CXR dataset. We compared the results generated by the three alignment methods proposed in R2GenGPT. For better illustration, the key medical information in the reports are highlighted using different colors.	131

List of Tables

3.1 Comparison on IU-Xray (upper part) and COV-CTR datasets (lower part). † indicates the results are quoted from the published literature. Specifically, we quote the results from (Li <i>et al.</i> , 2019a) for IU-xray, and from (Li <i>et al.</i> , 2023b) for COV-CTR. For other methods in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method.	31
3.2 Ablation studies. “Baseline” refers to RG-branch only. “ITM”, “FLL”, “STL” and “RL” stand for “Image-Text Matching”, “Feature Level Loss” ($\mathcal{L}_{feat}$ in Eqn. 3.1), “Self-boosted Triplet Loss” (Eqn. 3.3), and “Reinforcement Learning”, respectively. B@4 is the BLEU score uses 4-grams.	33
3.3 Hyper-parameter study using IU-Xray ($\lambda_{ITM} = 1$) . B@4 is the BLEU score uses 4-grams.	34
4.1 Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from the published literature. Specifically, we quote the results of CoAtt (Jing <i>et al.</i> , 2018) and HGRG-Agent (Li <i>et al.</i> , 2018) for IU-Xray dataset from (Li <i>et al.</i> , 2019a) while other methods we cite results from their respective papers. For methods without dagger tags in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method.	54
4.2 Model running time and capacity trained on IU-Xray dataset.	58
4.3 Performance comparisons on COCO Karpathy test split.	59

4.4 Ablation studies. We use our RG-branch only with the traditional cross entropy-loss as our baseline model for ablation experiments, which is denoted as “RG-branch(CE only)”. Here “CE”, “ITM”, “FLL”, “STL”, “RL”, and “SGCL” stand for “Cross-entropy Loss”, “Image-Text Matching”, “Feature Level Loss” ($\mathcal{L}_{feat}$ in Eqn. 4.1), “Self-boosted Triplet Loss” (Eqn. 4.3), “Reinforcement Learning”, and “Sample Graph Consistency Loss”($\mathcal{L}_{graph}$ in Eqn. 4.1), respectively.	60
4.5 Top 10 Anatomy entities and Observation entities with the highest frequency in Radgraph (Jain <i>et al.</i> , 2021)	60
4.6 Hyper-parameter study using MIMIC-CXR (fix $\lambda_{ITM} = 1$ and $\lambda_{SGCL} = 0$)	61
4.7 Hyper-parameter study using MIMIC-CXR (fix $\lambda_{ITM} = 1$, $\lambda_{CE} = 2$ and $\lambda_{feat} = 10$)	61
4.8 Performance using different parameter k on MIMIC-CXR dataset.	62
4.9 Performance using different alternate stride on MIMIC-CXR dataset.	63
4.10 Clinical accuracy evaluation. Keyword Accuracy (KA) is used to evaluate Clinical Correctness of the generated report, which is the ratio of the number of correctly generated keywords to the number of all keywords in the ground truth findings.	63
5.1 Performance comparison on MIMIC-CXR dataset.	74
6.1 Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from the published literature. Specifically, we quote the results from (Li <i>et al.</i> , 2019a) for IU-xray, and from (Chen <i>et al.</i> , 2020b) for MIMIC-CXR. For other methods in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method so that they are strictly comparable.	92

6.2 Ablation study on key components using IU-Xray. Here “baseline” refers to our pure transformer framework without the three criteria. “MLC”, “ITM”, “TW”, and “TmW” stand for “Multi-Label Classification”, “Image-text Matching”, “Term Weighting”, and “Temporal-Weighting”. BLue-4 is the BLUE score computed with 4-grams.	94
6.3 Ablation studies on hyper-parameters of Eqn. 6.9 using IU-Xray	95
6.4 Ablation studies on hyper-parameters using IU-Xray. We fix $\lambda_{RG} = 1$ and vary λ_{ITM} and λ_{MLC} for the investigation. The default values are: $\lambda_{RG} = 1$, $\lambda_{ITM} = 5$, and $\lambda_{MLC} = 2$.	96
7.1 Comparison of clinical efficacy metrics on the test set of the MIMIC-CXR dataset for measuring the accuracy of the description of clinical abnormalities.	112
7.2 Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from the published literature: the results of CoAtt (Jing <i>et al.</i> , 2018) and HGRG-Agent (Li <i>et al.</i> , 2018) on IU-Xray are quoted from (Li <i>et al.</i> , 2019a) while the other results are quoted from their respective papers. For the methods without †, their results are obtained by re-running the publicly released codebase (Li <i>et al.</i> , 2021) on these two datasets using the same training-test partition as our method.	113
7.3 Ablation studies. The “BASELINE” uses ViT encoder and bilinear-attention decoder with the traditional cross entropy-loss. Here “BE”, “ETs”, “OrL”, “EV” stand for “Bilinear-attention Encoder”, “Expert Tokens”, “Orthogonal Loss”, and “Expert Voting”. The number of expert tokens used in the experiments for the IU-xray and mimic-cxr datasets were 5 and 7, respectively. The average improvement over all NLG metrics compared to baseline is also presented in the “AVG. Δ ” column.	115

7.4 Comparison with the ensemble models on the IU-Xray dataset. The efficiency results refer to the total time required to complete the training using one RTX 3090.	117
8.1 Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from their respective papers. For the methods without †, their results are obtained by re-running the publicly released codebase (Li <i>et al.</i> , 2021) on these two datasets using the same training-test partition as our method.	128
8.2 Evaluation of Model Efficiency and Clinical Efficacy on MIMIC-CXR dataset.	130

Introduction

1.1 Background

The landscape of radiological imaging data is experiencing exponential growth that far surpasses the availability of trained readers, resulting in a significant and unsustainable surge in radiologists' workloads. This surge in both the volume and complexity of cases places big pressure on radiologists to interpret more studies within increasingly tight timeframes. Consequently, radiologists are faced with extended working hours and a heightened risk of reading fatigue, all of which significantly contribute to diagnostic errors. Notably, the situation is particularly precarious during on-call hours for emergency radiology studies (Katzman *et al.*, 2023). As a result, the demand for automated radiographic report generation has soared, as it promises to alleviate the burden on radiologists, mitigate diagnostic errors, and expedite the clinical workflow (Sim *et al.*, 2023).

Automated Radiographic Report Generation (R2Gen) stands as a sophisticated task in the field of artificial intelligence, aiming to generate coherent paragraphs that encapsulate the observations and findings in a given radiology image, as exemplified in Figure 1.1. These generated reports emulate the typical structure of radiological reports, which usually consist of "Findings" to detail the radiographic observations and "Impression" to summarize the clinical significance. While this structural format has a degree of standardization, it is essential to

recognize that report formats can indeed vary significantly across different regions and countries due to diverse clinical practices and documentation standards. Figure 1.2 showcases the diversity in report formats, using a Chinese medical report as an example. Notably, medical reports may exhibit differences in layout, terminology, and language, which have to be considered if global applicability is pursued. The primary focus of this thesis, however, is not to directly address this variation. Instead, our focus is directed towards enhancing the quality of medical report generation from radiographical images while specifically concentrating on reports in English with the standardized structure of 'Impression' and 'Findings' to control the complexity of our research endeavors. Despite this focus, it is noteworthy that the methodologies proposed in this thesis are language-agnostic and can be adapted to reports in any language. The adaptability is facilitated by employing language-specific tokenizers, which allows the core algorithm to remain unchanged.

This complex challenge is intrinsically tied to the broader area of research in image captioning, a subfield within computer vision. In image captioning, free-form text descriptions are generated to narrate the content of images. Research of image captioning falls in three categories: templated based, retrieval based, and generation based. Template-based approaches (Kisilev *et al.*, 2015b; Kisilev *et al.*, 2015a; Kisilev *et al.*, 2016) generate captions by filling a predefined template. Despite having the merit of simplicity, this approach generates rigid sentences with a fixed length, governed by human-crafted rules. Retrieval-based approaches (Devlin *et al.*, 2015) assume that similar images should have similar captions and generated the caption of an given image by first finding visually similar images with their captions and then retrieving from these existing captions. These approaches have the limitation to generate image-specific and semantically correct captions. Current image captioning research is dominated by generation based methods using deep neural network. A basic deep neural

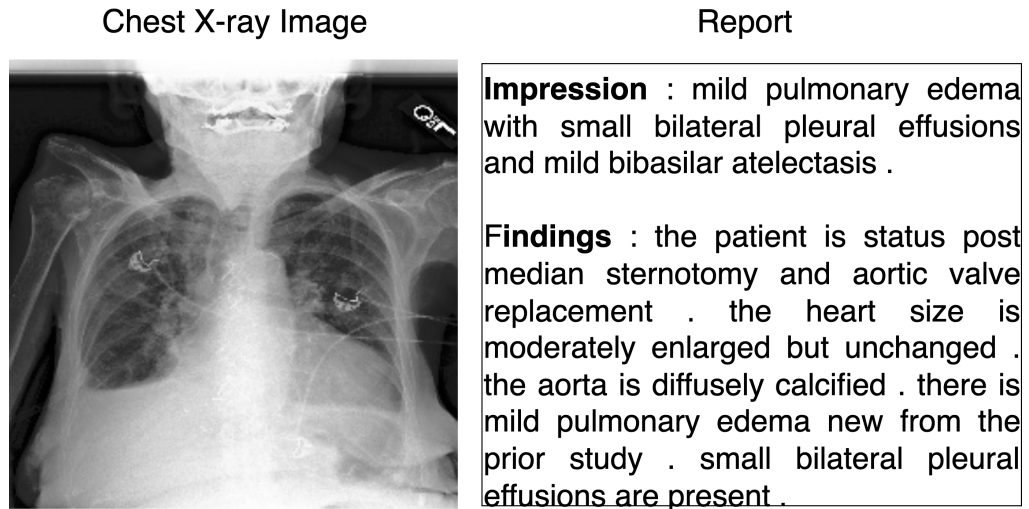


Figure 1.1. An example of Chest X-ray data used in this thesis.

network model generating image captions follows the encoder-decoder structure (Vinyals *et al.*, 2015), where the visual encoder extracts the visual features from images and the text decoder converts the visual features to text output. The development along this research line focuses on advancing encoder and decoder, respectively, by employing carefully-designed attention mechanisms (Lu *et al.*, 2017; Anderson *et al.*, 2018; Rennie *et al.*, 2017), relationship among image regions (Yao *et al.*, 2018; Yang *et al.*, 2019), improved language models (Devlin *et al.*, 2019; Cornia *et al.*, 2020; Pan *et al.*, 2020), or reinforcement learning on language metric (Rennie *et al.*, 2017; Liu *et al.*, 2017; Pasunuru and Bansal, 2017). A detailed review about the deep learning based image captioning methods could be found in Section 2.1.

However, compared to the image captioning task, radiographic report generation presents several unique challenges. The first challenge lies in **long text generation**. Unlike the image captioning task which generates a single sentence description, medical report generation requires detailed and coherent paragraph-long descriptions. This requires the model to have a robust capacity for learning long-range dependencies. The second challenge lies in **fine-grained image**

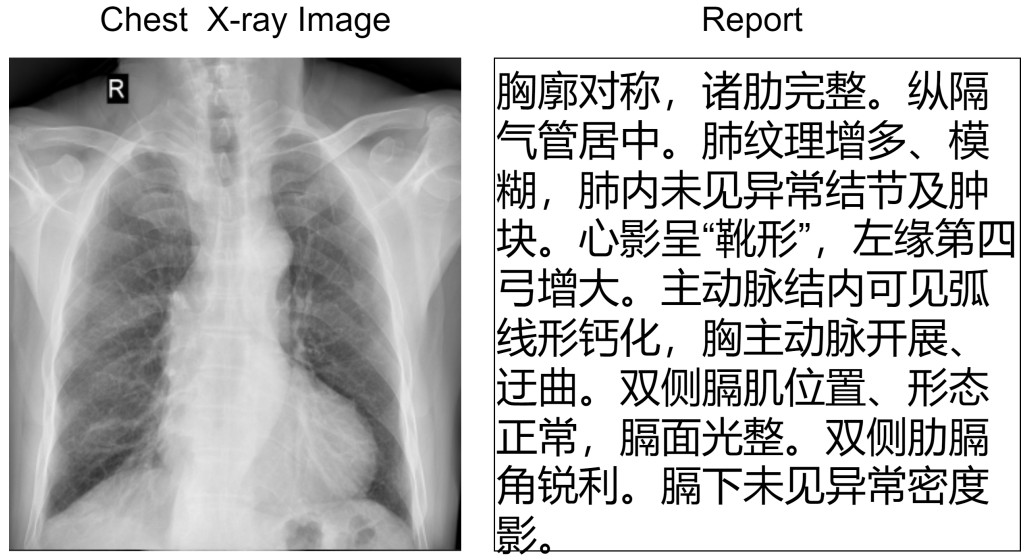


Figure 1.2. An example of a Chest X-ray alongside a corresponding radiological report in Chinese.

distinctions. Radiological images like chest X-rays often possess striking similarities, making it difficult for the model to identify nuanced differences among them. This limitation hampers the model’s capacity to accurately interpret and report the underlying medical conditions. The third challenge lies in the **bias in visual and textual data**. Due to an over-representation of normal samples in the training data, the model’s learning process was biased towards these samples, limiting its ability to effectively detect abnormalities and anomalies within the dataset.

Numerous R2Gen models have been proposed to address these challenges. Chen et al. (Chen et al., 2020b) introduced additional memory modules to store past information, which can be utilized during the decoding process to improve long-text generation performance. Some work utilize external knowledge information, with the most representative approach being the integration of knowledge graphs (Li et al., 2019a; Zhang et al., 2020; Liu et al., 2021a; Yang et al., 2021b; Li et al., 2023a; Huang et al., 2023). Zhang et al. (Zhang et al., 2020)

and Liu et al. (Liu *et al.*, 2021a) combined pre-constructed graphs representing relationships between diseases and organs using graph neural networks, enabling more effective feature learning for abnormalities. Li et al. (Li *et al.*, 2023a) developed a dynamic approach that updates the graph with new knowledge in real-time. Huang et al. (Huang *et al.*, 2023) incorporated knowledge from a symptom graph into the decoding stage using an injected knowledge distiller. Apart from knowledge graphs, another method for integrating external knowledge involves incorporating semantic information to assist report generation through multi-task learning, such as multi-label classification (Jing *et al.*, 2018; Jin *et al.*, 2023; Tanida *et al.*, 2023). Jin et al. (Jin *et al.*, 2023) make use of the diagnostic results from the classification via prompts to explicitly guide the generation process. A detailed review could be found in Section 2.2.

Despite significant advancements in the field, there is still considerable room for improvement in radiographic report generation. In this thesis, aiming at the above-mentioned challenges, we design and propose multiple network architectures and improvement strategies. These methods will be elaborated in detail from Chapter 3 to Chapter 8.

1.2 Outline

This thesis is organized as follows:

- In Chapter 1, we introduce the importance and challenges of automated radiographic report generation. We outline our models and main contributions, setting the stage for the novel methodologies presented in the subsequent chapters.
- In Chapter 2, we conduct a comprehensive Literature Review, covering key advancements in image captioning, radiology report generation, Transformer architectures, image-text matching techniques, and

the emergence of large language models. This review sets the stage for understanding the current landscape and gaps in the relevant fields.

- In Chapter 3, we present our self-boosting framework designed to improve the quality of radiographic report generation. We introduce a dual-branch network model consisting of the main task of report generation and an auxiliary task of image-text matching. The chapter explains how these two branches cooperatively influence each other to enhance the report’s quality. The loss function, which goes beyond word similarity to consider feature-level similarity, is also discussed. Finally, we provide experimental results to demonstrate the method’s effectiveness against state-of-the-art methods in both image captioning and medical report generation.
- In Chapter 4, we extend the self-boosting framework by integrating two novel strategies aimed at achieving better correlation between image and text features. The first strategy involves the auxiliary task of image-text matching (ITM), while the second strategy focuses on aligning the sample spaces of images and reports. We elucidate how these strategies make our framework more robust, and present additional experimental results on medical report generation benchmarks, including the largest dataset MIMIC-CXR.
- In Chapter 5, we address the difficulty in capturing fine-grained visual differences in radiographic images and the inadequacy of existing methods in providing effective medical semantic features. We propose a memory-augmented sparse attention block and a Medical Concepts Generation Network (MCGN) to meet these challenges. Experimental validation on the MIMIC-CXR dataset shows the superiority of our method over existing state-of-the-art solutions.
- In Chapter 6, we propose a pure transformer-based framework that enhances visual-textual alignment and multi-label diagnostic classification for better report generation. We discuss novel mechanisms for

embedding auxiliary objectives and term-weighting schemes, concluding with promising results on benchmark datasets.

- In Chapter 7, we explore the concept of "multi-expert joint diagnosis," introducing METransformer. This method utilizes multiple learnable "expert" tokens for both the encoder and decoder stages in a transformer-based architecture. We demonstrate how this approach outperforms single-expert models on various benchmarks.
- In Chapter 8, we investigate the potential of Large Language Models (LLMs) for Radiology Report Generation (R2Gen). Introducing R2GenGPT, we discuss how we align visual features with word embeddings from LLMs. We showcase the benefits of this approach in terms of training efficiency and state-of-the-art performance.
- In Chapter 9, we conclude the thesis and discuss the future outlook.

1.3 Contributions

The main contributions of the thesis are summarized as follows.

- In Chapter 3, we develop a self-boosting framework integrating an auxiliary task of image-text matching with medical report generation. This tightly-coupled, mutually reinforcing approach improves the fine-grained diagnostic capabilities of generated reports. The method introduces a new loss term for evaluating feature similarity between generated and ground-truth reports.
- In Chapter 4, we extend the self-boosting framework to focus on the fine-grained characteristics of diagnostic captioning. We introduce a sample-graph consistency loss to enforce consistent visual and textural feature embeddings.
- In Chapter 5, we introduce a bilinear-pooling-assisted sparse attention block embedded in a transformer network to enhance the capture of

fine-grained visual differences in radiographic images. We also propose a medical concepts generation network to enrich semantic information for radiographic report generation.

- In Chapter 6, we pioneer the use of a pure transformer model for medical report generation, breaking away from the conventional CNN-based methods. We incorporate a multi-criteria supervision scheme that includes an image-text matching loss for semantic coherence and a term-weighting scheme coupled with multi-label classification for capturing critical diagnostic information. This innovative approach maintains simplicity without altering the transformer structure.
- In Chapter 7, we introduce METransformer, a novel "multi-expert joint diagnosis" framework for radiology report generation. Through the use of learnable expert tokens and both linear and non-linear attentions, the model learns complementary representations. This carefully designed ensemble-like approach outperforms traditional ensemble methods in both efficiency and performance.
- In Chapter 8, we present R2GenGPT, a novel framework utilizing pre-trained large language models (LLMs) for the Radiology report generation task. The framework introduces three methods—shallow alignment, delta alignment, and deep alignment—for effectively integrating visual information into LLMs.
- Across all chapters, each of the proposed frameworks and methods has been extensively validated and has set new performance standards on various widely-used benchmarks, including but not limited to the MIMIC-CXR and IU-Xray datasets. They consistently demonstrate superior performance and hold promising potential for further advancements in the field of Radiology report generation.

Literature Review

2.1 Image Captioning

Natural image captioning task aims at automatically generating sentence description for the given image. A broad collection of methods was proposed in the last few years (Karpathy and Fei-Fei, 2015; Vinyals *et al.*, 2015; Donahue *et al.*, 2015; Xu *et al.*, 2015; You *et al.*, 2016; Yang *et al.*, 2016; Lu *et al.*, 2017; Rennie *et al.*, 2017; Liu *et al.*, 2017; Anderson *et al.*, 2018; Cornia *et al.*, 2020; Zhou *et al.*, 2020; Pan *et al.*, 2020). Most of them adopt the encoder-decoder architecture, employing CNN as the encoder to extract the visual feature of the image and RNN as the decoder to produce image description. Among these methods, *Show-Tell* (Vinyals *et al.*, 2015) is the most canonical neural image captioning model, which provides an end-to-end framework by feeding the image features extracted by CNN as the input of the LSTM (Hochreiter and Schmidhuber, 1997) to produce image captions. On the basis of this framework, inspired by human brain’s attention, several methods have proposed the attention mechanism (Xu *et al.*, 2015; You *et al.*, 2016; Yang *et al.*, 2016; Lu *et al.*, 2017), allowing the model to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. For example, in (Lu *et al.*, 2017) an adaptive attention model was proposed to automatically switch the focus between visual signals and the language model; in (Anderson *et al.*, 2018) a combined bottom-up and top-down attention mechanism calculated attention at the level of objects and salient regions. Also, in a very recent work in (Zhou *et al.*, 2020), it was proposed to learn attention from an additional

image-text matching task and use it to regularize the image captioning task. Although this work is close to our work in utilizing image-text matching for image captioning, they are substantially different as explained in Section 3.2.5. The superiority of our method over (Zhou *et al.*, 2020) has also been verified in our experiment. In addition to developing attention models, other image captioning works also explore how to advance the image encoder and text decoder. To improve the image encoder, the methods (Yao *et al.*, 2018; Yang *et al.*, 2019) explicitly considered the visual relationship of the detected image regions by constructing graphical convolutional network (GCN) or scene graph. To improve the text decoder, in (Krause *et al.*, 2017) a hierarchically structured RNN was developed to cater for paragraph generation and the work in (Pan *et al.*, 2020) proposed a X-Linear attention block and integrated it into image captioning structures to selectively capitalize on visual information. Recently, Transformer model (Vaswani *et al.*, 2017) has been proposed to improve the limited representation power of RNN, and has been used in image captioning tasks to replace RNN as the text decoder (Cornia *et al.*, 2020). Moreover, recently Reinforcement Learning has also been introduced to train the non-differentiable captioning metrics (Rennie *et al.*, 2017; Liu *et al.*, 2017; Pasunuru and Bansal, 2017) for improvement.

Despite recent advancements in image captioning, a notable decline in performance is often observed when directly applying these techniques to the generation of medical reports. This is primarily due to the unique challenges inherent in radiology report generation, as opposed to general image captioning, which we have discussed in Section 1.1. In the following section, we will explore how current approaches in radiology report generation have been refined to effectively address these challenges.

2.2 Radiology report generation

Radiology Expert Reporting and Interpretation Process The creation of a radiology report is the culmination of a nuanced interpretive process, blending scientific acumen with communicative precision (Waite *et al.*, 2019). When a radiologist examines a radiographic image, the task involves an intricate analysis that goes beyond mere observation. Each image is carefully scrutinized; the radiologist must discern subtle variations in tissue density, structure alignment, and other radiographic signs that may indicate underlying pathology (Hosny *et al.*, 2018).

To systematically construct the report, the radiologist first documents the 'Findings'. This involves a detailed description of each radiographic feature noted, which may include the size, shape, location, and any changes from previous studies. The language used is specific and standardized, minimizing ambiguity and ensuring consistency across the medical field.

Following the objective documentation of findings, the radiologist engages in a higher level of cognitive synthesis to craft the 'Impression' section. This is where interpretative skill is paramount, as the radiologist must integrate the findings with clinical knowledge to form a diagnostic hypothesis. The impression provides a concise summary that highlights the most clinically significant aspects of the findings, offers a possible diagnosis, and sometimes includes a differential diagnosis when multiple conditions could explain the radiographic findings (Alexander *et al.*, 2020; Hartung *et al.*, 2020).

Furthermore, the impression may offer recommendations for further investigations, such as additional imaging studies, laboratory tests, or clinical follow-up, providing a clear direction for subsequent patient care. These recommendations are formulated based on the radiologist's experience, the current clinical scenario, and the latest medical guidelines and research findings.

AI-Driven Advances in Radiology Report Generation Radiology report generation has witnessed significant advancements propelled by novel model architectures and the infusion of specialized domain knowledge. One of the mainstays in this domain has been the hierarchical Long Short-Term Memory (LSTM) networks (Jing *et al.*, 2018; Yuan *et al.*, 2019; Yin *et al.*, 2019; Wang *et al.*, 2021). For instance, Jing *et al.* (Jing *et al.*, 2018) innovated a hierarchical model integrated with a co-attention mechanism and multi-task learning to automate keyword prediction, facilitating the generation of longer paragraphs. This strand of research is paralleled by the work of Yin *et al.* (Yin *et al.*, 2019), who introduced a topic-matching mechanism for better semantic alignment between sentences. Contributions from Xue *et al.* (Xue *et al.*, 2018; Xue and Huang, 2019) further complement the hierarchical LSTM architecture. Xue *et al.* presented a model involving a generative sentence model and a recurrent paragraph generative model to dynamically produce text (Xue *et al.*, 2018; Xue and Huang, 2019).

Moving beyond LSTM-based architectures, an emerging trend is the adoption of Transformer models as text decoders (Chen *et al.*, 2020b; Liu *et al.*, 2021a; You *et al.*, 2021). Notably, Chen *et al.* (Chen *et al.*, 2020b) designed a memory-driven Transformer to capture key information during the report generation process.

A parallel line of research has aimed to integrate domain knowledge, particularly through knowledge graphs (Zhang *et al.*, 2020; Yang *et al.*, 2021a; Yang *et al.*, 2021b; Liu *et al.*, 2021b; Li *et al.*, 2019a; Liu *et al.*, 2021a; Li *et al.*, 2023a; Huang *et al.*, 2023). For example, Yang *et al.* (Yang *et al.*, 2021a) and Liu *et al.* (Liu *et al.*, 2021b) dynamically update medical knowledge bases to augment visual feature extraction. Similarly, Zhang *et al.* (Zhang *et al.*, 2020) and Liu *et al.* (Liu *et al.*, 2021a) employ pre-constructed graphs to represent relationships between diseases and organs, facilitating more effective feature learning for identifying abnormalities.

Furthermore, multi-task learning strategies like multi-label classification have also been introduced to guide report generation (Jing *et al.*, 2018; Wang *et al.*, 2022c; Wang *et al.*, 2022b; Jin *et al.*, 2023; Tanida *et al.*, 2023). Jing *et al.* (Jing *et al.*, 2018) initially connected an image encoder’s output to a disease classification network, and then used the prediction results as the LSTM’s hidden states to guide the generation of the report. Wang *et al.* (Wang *et al.*, 2022c) even mined high-frequency medical terms from RadGraph (Jain *et al.*, 2021) to train a multi-label classification network, the outputs of which were then incorporated into the decoding process.

The collective thrust of these developments showcases a multidisciplinary effort to push the boundaries of automated medical report generation, employing both innovative machine learning techniques and domain-specific knowledge for improved performance and reliability.

2.3 Transformers

Transformers have achieved state-of-the-art (SOTA) results not only in various natural language processing tasks but also in computer vision (Dosovitskiy *et al.*, 2021). The architecture has been particularly effective in image captioning, with seminal works employing the model to exploit both semantic and visual information (Yu *et al.*, 2019; Li *et al.*, 2019b; He *et al.*, 2020; Herdade *et al.*, 2019; Cornia *et al.*, 2020). Marcella *et al.*’s mesh-connected transformer framework (Cornia *et al.*, 2020) stands out for its unique use of each encoder layer’s output as input for each decoder layer. Similarly, an entangled attention module was proposed in (Li *et al.*, 2019b) to allow the transformer to consider both visual and semantic information concurrently. Despite these advancements in the general image captioning domain, the application of transformers in medical report generation is still nascent but promising (Chen *et al.*, 2020b; Liu *et al.*, 2021a). Chen *et al.* (Chen *et al.*, 2020b) introduced a memory-driven transformer

specifically for Chest X-ray report generation, incorporating a relational memory module to recall information from previous stages of the generation process. Another recent work (Liu *et al.*, 2021a) innovatively constructed three separate transformer networks to explore and distill both posterior and prior knowledge for medical report generation. Interestingly, although most image captioning approaches still rely on convolutional neural networks (CNNs) to extract visual features (Yu *et al.*, 2019; Li *et al.*, 2019b; He *et al.*, 2020; Herdade *et al.*, 2019; Cornia *et al.*, 2020), the work by Dosovitskiy *et al.* (Dosovitskiy *et al.*, 2021) demonstrated that a pure transformer could attain SOTA performance in image classification tasks, opening up new avenues for exploration in vision-to-language tasks, particularly for medical report generation.

2.4 Image-Text Matching

Image-text matching aims to assess the visual-semantic similarity between textual and visual elements, and has proven integral in various applications including image captioning and medical report generation. **This section is particularly pertinent to Chapter 3, Chapter 4, and Chapter 6 of this thesis, where we leverage image-text matching networks to enhance the performance of radiology report generation.** The mainstay approach for this task involves projecting both image and text features into a unified semantic space, where similarity metrics such as cosine distance are applied (Kiros *et al.*, 2014; Wang *et al.*, 2016; Faghri *et al.*, 2018). Methods in this field predominantly fall under two categories: those that employ global representations (Kiros *et al.*, 2014; Wang *et al.*, 2016; Faghri *et al.*, 2018) and those that focus on local, attention-based representations like SCAN (Lee *et al.*, 2018; Liu *et al.*, 2019; Wang *et al.*, 2020). SCAN’s attention mechanism, in particular, has been adapted to enrich image captioning tasks (Zhou *et al.*, 2020). Unlike the approach in (Zhou *et al.*, 2020), where the two tasks are trained independently, in the Chapter 3 and Chapter 4, our work leverages a unified network model for both image-text

matching and report generation, enabling a more synergistic learning of textual and visual features. Recently, transformers have been introduced into the image-text matching domain (Chen *et al.*, 2020a), using unified transformer encoders to process both image and text features. However, our method in Chapter 6 takes it a step further by employing both the encoder and decoder aspects of the transformer, facilitating a nuanced approach that addresses the complexities of both image-text matching and medical report generation within a singular framework.

2.5 Large language Models

Recently, there has been a surge of interest in Large Language Models (LLMs) due to their superior efficacy in a wide array of Natural Language Processing (NLP) tasks. **This section is especially relevant to Chapter 8, where we investigate the use of Large Language Models in the generation of radiology reports.** This began with transformer models like BERT (Su *et al.*, 2020), GPT (Radford *et al.*, 2018), and T5 (Raffel *et al.*, 2020), each designed with distinct pre-training objectives. The introduction of GPT-3 (Brown *et al.*, 2020) marked a significant shift, demonstrating the model’s impressive zero-shot generalization capabilities, owed to a scaled-up parameter and data volume, which allowed it to excel in tasks it had not previously encountered. This catalyzed the development of several LLMs such as OPT (Zhang *et al.*, 2022), BLOOM (Scao *et al.*, 2022), PaLM (Chowdhery *et al.*, 2022), and LLaMA (Touvron *et al.*, 2023), heralding the triumph of LLMs. In a parallel endeavor, Ouyang *et al.* (Ouyang *et al.*, 2022) introduced InstructGPT, which brought human instruction and feedback into alignment with GPT-3. These advancements have been leveraged by applications like ChatGPT, which enables human-like dialogue interactions by responding to a vast spectrum of complex and nuanced questions and instructions.

A Self-boosting Framework for Automated Radiographic Report Generation

Automated radiographic report generation is a challenging task since it requires to generate paragraphs describing fine-grained visual differences of cases, especially for those between the diseased and the healthy. Existing image captioning methods commonly target at generic images, and lack mechanism to meet this requirement. To bridge this gap, in this chapter, we propose a self-boosting framework that improves radiographic report generation based on the cooperation of the main task of report generation and an auxiliary task of image-text matching. The two tasks are built as the two branches of a network model and influence each other in a cooperative way. On one hand, the image-text matching branch helps to learn highly text-correlated visual features for the report generation branch to output high quality reports. On the other hand, the improved reports produced by the report generation branch provide additional harder samples for the image-text matching branch and enforce the latter to improve itself by learning better visual and text feature representations. This, in turn, helps improve the report generation branch again. These two branches are jointly trained to help improve each other iteratively and progressively, so that the whole model is self-boosted without requiring external resources. Additionally, in the loss function, our model evaluates the quality of the generated reports not only on the word similarity as common approaches do (via minimizing a cross-entropy loss), but also on the feature similarity at high-level,

which is provided by the report encoder of the image-text matching branch. Experimental results demonstrate the effectiveness of our method on two public datasets, showing its superior performance over multiple state-of-the-art image captioning and medical report generation methods.

3.1 Introduction

The rising demand for radiologists to interpret an ever-increasing volume of medical images has made automated medical report generation an essential area of research. While traditional image captioning techniques offer some promise, when directly applying image captioners for medical report generation, there is often a visible performance decline. This is because compared with generic images, radiographic images are more similar to each other and fine-grained visual differences, such as the findings of clinic importance, need to be narrated. This requires further tightening the visual and text representations. In addition to linking image and text by attention mechanism, some medical report generation methods (Jing *et al.*, 2018; Yin *et al.*, 2019) additionally train classifiers to learn visual representations by predicting tags of medical reports or disease-class labels. This often requires additional annotations and external medical datasets or knowledge, which are task-specific and often unavailable, limiting the generalization of these methods. Moreover, either tags or disease classes only sparsely cover the reports' information, leading to relatively loosely correlated visual and text features.

To bridge this gap, in this chapter, we propose a self-boosting framework to promote radiographic report generation. It utilizes an auxiliary task to predict the match of an image-report pair (i.e., image-text matching) by explicitly learning strongly correlated visual and text features. These features could better serve the fine-grained recognition task in radiographic report generation. More importantly, the auxiliary image-text matching task is deeply coupled with the

main report generation task through our proposed self-boosting framework so that these two tasks could mutually and progressively boost each other via the cooperative interactions between them. Specifically, these two tasks are built as two branches of a neural network model and jointly trained. During the training iterations, the image-text matching branch provides better features to the report generation branch to generate ground-truth-like reports, and these improved reports, in turn, serve as additional harder samples to push the image-text matching branch to become stronger by learning better features. In this way, the whole network is gradually self-boosted and improves the ultimate report generation performance.

Our main contributions are summarized as follows.

First, we utilize an auxiliary task of image-text matching to help learn text-correlated visual features that could help capture the fine-grained visual differences of diagnostic importance. This improves the main task of report generation without requiring additional annotations, external medical datasets, or external task-specific knowledge.

Secondly and more importantly, we propose a self-boosting framework that leverages the cooperation between image-text matching and report generation to mutually boost each other progressively. These two tasks are tightly coupled and jointly trained, while the stronger features learned by image-text matching help improve report generation, and the improved reports, as additional harder samples, in turn enforce image-text matching to continue improving feature learning so that the finer mismatch between an image and its generated report could be identified.

Third, additionally utilizing image-text matching also allows us to learn an effective text feature extractor, which is used to evaluate the feature similarity between the generated reports and the ground-truth, providing a new loss term to further promote the report generation.

Fourth, our approach shows promising performance on two benchmarks, generating reports from both classic Chest X-ray images and CT images with COVID-19.

3.2 Method

In this section, we first give an overview of the proposed model. Then two important branches of our model, i.e., image-text matching and medical report generation, are presented. Following that, we describe how these two branches achieve joint training and mutual improvement.

3.2.1 Overview

A typical radiographic report consists of a conclusive “impression” part and a descriptive “finding” part. Taking a medical image as input, our model generates a long-paragraph diagnostic report with both parts. As mentioned in Section 3.1, our model improves report generation by taking advantage of the interactions between the main task of report generation and the auxiliary task of image-text matching. Our model is trained in a self-boosting manner.

Figure 3.1 provides an overview about our framework. In general, our model consists of two entangled branches: medical report generation (main task) and image-text matching (auxiliary task). As for the main task, our report generation branch (RG-branch) follows the encoder-decoder architecture. The visual encoder detects image regions with an unsupervised method, extracts regional features using CNN, and refines these visual features based on the regional relationships learnt through self-attention. The text decoder takes the refined visual features as input, passes them through a hierarchical LSTM model catering for both the topic-level and word-level decoding, and generates the final paragraphical report. As for the auxiliary task, taking the image-report pairs as input, our

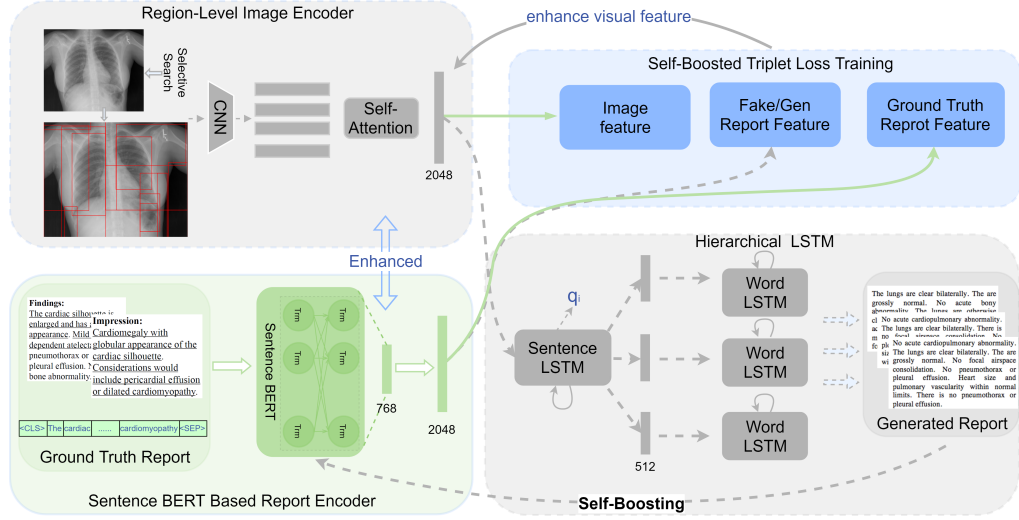


Figure 3.1. An overview of the proposed framework comprising of a medical report generation branch (RG-branch, the gray part) and an auxiliary image-text matching branch (ITM-branch, the green part). RG-branch consists of two components: a region-level image encoder (shared with ITM-branch) to extract visual features from medical images, and a hierarchical LSTM to generate diagnose reports. ITM-branch consists of an image encoder (shared with RG-branch) and a report encoder based on Sentence-BERT. It compares the similarity of visual and text features for match. The two branches are designed to have cooperative interactions during training. ITM-branch provides RG-branch text-correlated visual features for report generation and its report encoder for measuring the feature level loss of the generated reports. In turn, with the improved generated reports, RG-branch supplies ITM-branch harder and harder samples to force the latter to evolve for better feature learning. In this way, the whole model is self-boosted.

image-text matching branch (ITM-branch) consists of an image encoder and a report encoder, learning the visual and the text features, respectively, in order to predict if the input image and report match each other.

The two branches interact with each other in three ways. First, they share the visual encoder, so that the text-correlated visual features learned by ITM-branch could also be utilized to improve the RG-branch to generate high-quality reports. Second, the report encoder learnt by the ITM-branch is also utilized by the RG-branch to evaluate and minimize the feature-level loss between the generated and the ground-truth reports. Third, during the training, the improved reports generated by RG-branch are passed to the ITM-branch as additional harder samples, which enforces the ITM-branch to improve itself by enhancing the feature

learning. These three interactions last for the training course, making the whole model gradually boost itself. At inference stage, only RG-branch is used, which receives a test image and generates the corresponding report.

3.2.2 Report Generation (RG) Branch

The ultimate target of our model is to generate ground-truth-like diagnostic reports from radiographic images. Therefore, report generation is our main task. Like most image-captioning methods, our RG-branch consists of a visual encoder and a successive text decoder, with enhanced characteristics. They are elaborated as follows.

Visual encoder Although some generic image captioners use Faster R-CNN (Ren *et al.*, 2017a) to encode image regions at object-level, existing medical report generation methods extract visual features from the whole image rather than image regions, due to the lack of supervised information. Differently, given an input image of size $3 \times H \times W$, we employ the selective search algorithm to generate region proposals unsupervisely, and then refine these region proposals by: 1) excluding regions smaller than 2000 pixels; 2) excluding regions where the background ratio exceeds 70%. This leads to roughly $M = 30$ regions for one image. After that, we adopt ResNet-101 to produce a set of vectors $v_1, \dots, v_M \in \mathbb{R}^D$ with a dimension $D = 2048$ for each region. Compared with image-level features, using regional features can “look closer” to the image, thus better suits the fine-grained pattern description. It is found that modeling region relationship is beneficial for visual representation. Unlike some image-captioning methods (Yao *et al.*, 2018; Yang *et al.*, 2019) that rely on the Visual Genome dataset to learn GCN or scene graph for generic images, we employ self-attention (Vaswani *et al.*, 2017) to learn the relationships among regions and transform regional visual features $\mathbf{V}$ into relationship-enhanced ones $\mathbf{V}_h$ for refinement. These relationship-enhanced regional features are then passed

through a region pooling layer to generate the pooled vector $\mathbf{v}_p$. We further input the pooled vector $\mathbf{v}_p$ through a 2048-dimensional fully-connected layer to produce the final representation: $\mathbf{I}_f = \mathbf{W}_f \mathbf{v}_p + \mathbf{b}_f$. The output of the visual encoder $\mathbf{I}_f$ is then sent to the text decoder in RG-branch for report generation, as well as to ITM-branch for predicting image-text matching.

Text decoder To generate paragraphical report, we adopt a hierarchical structured LSTM, which includes a sentence LSTM and a word LSTM (Krause *et al.*, 2017). The sentence LSTM takes the visual features as input, produces a *topic vector* for each sentence in the report paragraph, and determines when to stop generating the topic vector. Those topic vectors are then fed into the word LSTM to produce the sentences. More details are as follows.

The sentence LSTM is a single layer LSTM with a hidden size $H = 512$. It takes the image feature $\mathbf{I}_f$ as input, and in turn generates a sequence of hidden states $\mathbf{h}_1 \cdots \mathbf{h}_N \in \mathbb{R}^H$, one for each sentence in the paragraph: $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{I}_f, (\mathbf{h}_i, \mathbf{c}_i))$, $i \in \{0, \cdots, N-1\}$, where $\mathbf{c}_i$ denotes the memory cells of LSTM, and N is the number of sentences in the paragraph. Based on the hidden state $\mathbf{h}_i$, a topic vector $\mathbf{t}_i \in \mathbb{R}^P$ is obtained as: $\mathbf{t}_i = \tanh(\mathbf{W}_t \mathbf{h}_i + \mathbf{b}_t)$ $i \in \{1, \cdots, N\}$. In addition, the hidden state $\mathbf{h}_i$ is also mapped from H dimension to two dimension by a linear projection to produce a probability q_i through a sigmoid function, determining whether the i th sentence is the last sentence in the paragraph. q_i can be learnt by minimizing the cross entropy loss for a binary classification of being continued or not. In our work, when the “impression” part is available in the reports, we further use this information to guide the generation of the topic vectors, since “impression” implies a global topic about the report. Specifically, a single layer LSTM is first trained to generate “impression”, and then the attained weights are utilized to initialize the weights of the sentence LSTM.

The word LSTM is also a single layer LSTM with hidden size $H = 512$ and the initial hidden and cell states of the word LSTM are set to zero. It takes the topic vectors produced by the sentence LSTM as the input and generates detailed sentences. Given the i th topic t_i , the hidden and cell states of the word LSTM are updated by $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{s}_{j-1}^{(i)}, (\mathbf{h}_i, \mathbf{c}_i))$, where $j \in \{0, \dots, M+2\}$, and M is the number of word in i th sentence. It should be noted, when $j = 0$, $\mathbf{s}_{-1}^i$ represents the topic vector $\mathbf{t}_i$; when $j > 0$, $\mathbf{s}_j^i$ represents the word embedding of the j th word in the i th sentence. In particular, $\mathbf{s}_0$ and $\mathbf{s}_{M+1}$ represent the word embedding of the special token $\langle start \rangle$ and $\langle end \rangle$ respectively. After obtaining the hidden state of word LSTM, the probability $p_{ij} = \text{Softmax}(\mathbf{W}_p \mathbf{h}_j + \mathbf{b}_p)$ can be learnt, predicting the probability of j th word in i th sentence, where $i \in \{1, \dots, M\}, j \in \{1, \dots, N\}$.

Loss function Most image captioning and medical report generation methods construct the loss functions evaluating the difference between the generated and the ground-truth reports at the word-level (cross-entropy based loss).

In this chapter, we further argue that a generated report with high-quality should also be similar to the ground-truth at the semantic feature levels. That is, they should stay similar after they are passed through a text feature extractor producing features corresponding to high-level semantics. In our work, such a text feature extractor could be readily achieved by utilizing the report encoder learnt in the auxiliary ITM-branch, which is well trained to effectively encode the most critical text features correlated to the visual contents. Specifically, given K reports, the loss function used in our RG-branch consists of two terms as follows:

$$\mathcal{L}_{RG} = \lambda_{CE} \sum_{k=1}^K \mathcal{L}_{CE}^{(k)} + \lambda_{feat} \sum_{k=1}^K \mathcal{L}_{feat}^{(k)}, \quad (3.1)$$

where λ_{CE} and λ_{feat} are hyper-parameters balancing the two terms.

The first loss term $\mathcal{L}_{CE}^{(k)}$ is the cross-entropy loss used to measure the word-level difference for the k th report. Since we use a hierarchical LSTM as the

text decoder, the $\mathcal{L}_{CE}^{(k)}$ is the sum of the cross-entropy loss $\mathcal{L}_S^{(k)}$ of the sentence LSTM and the cross-entropy loss $\mathcal{L}_W^{(k)}$ of the word LSTM. The loss $\mathcal{L}_S^{(k)} = -\sum_{i=0}^N \left[l_i^{(k)} \cdot \log(q_i^{(k)}) + (1 - l_i^{(k)}) \cdot \log(1 - q_i^{(k)}) \right]$, where $q_i^{(k)}$ and $l_i^{(k)}$ are the predicted and the ground-truth labels, respectively, indicating if the i th sentence is the last sentence in the k th report. The loss $\mathcal{L}_W^{(k)} = -\sum_{i=1}^N \sum_{j=1}^M y_{ij}^{(k)} \log(p_{ij}^{(k)})$, where $y_{ij}^{(k)}$ refers to the ground-truth and $p_{ij}^{(k)}$ refers to the prediction of the j th word in the i th sentence within the k th report.

The second loss term $\mathcal{L}_{feat}^{(k)} = \|\Phi^{ITM-R}(\hat{\mathbf{T}}^{(k)}) - \Phi^{ITM-R}(\mathbf{T}^{(k)})\|_2$ measures the difference at the feature level, where $\hat{\mathbf{T}}^{(k)}$ and $\mathbf{T}^{(k)}$ are the predicted paragraph and the ground-truth for the k th report, respectively. Here the mapping $\Phi^{ITM-R}(\cdot)$ denotes the embedding function of the report encoder in ITM-branch (see Section 3.2.3 for more details), which is highly nonlinear.

In addition, following literature (Rennie *et al.*, 2017), we also apply reinforcement learning with CIDEr as the reward to further boost the performance of our RG-branch.

3.2.3 Image-Text Matching (ITM) Branch

To promote the main task of report generation, we also introduce an auxiliary task of image-text matching to learn highly correlated visual and text features. Our ITM-branch takes an image-report pair as input and tells whether the image and the report match. It consists of an image encoder and a report encoder, detailed as follows.

Image encoder Our ITM-branch shares its image encoder with our RG-branch. As mentioned above, the image encoder extracts visual features from regions, refines these features with regional relationship, and outputs the image representation on top of an additional region pooling. It is worth mentioning that, sharing the image encoder between ITM-branch and RG-branch enables our

model to both learn image-text correlated features and fine-tune these features for the main report generation task.

Report encoder We considered BERT’s (Devlin *et al.*, 2019) model as the report encoder, which is a multi-layer bidirectional Transformer (Vaswani *et al.*, 2017) and has set a new state-of-the-art performance on various NLP tasks. However, we do not directly use the vanilla BERT. As mentioned, the report encoder in ITM-branch is also employed by RG-branch to compare the feature similarity between the generated and the ground-truth reports. Taking this into account, we adapt a

Sentence-BERT (Reimers and Gurevych, 2019) model to constructing our report encoder, because Sentence-BERT employed a siamese and triplet network structure (Schroff *et al.*, 2015) for BERT, making it more suitable to derive semantically meaningful sentence embeddings.

Sentence-BERT has been well pre-trained on two large broad-coverage corpus (Bowman *et al.*, 2015; Williams *et al.*, 2018) for various textual similarity tasks. Moreover, on top of the pretrained Sentence-BERT, we stack another fully connected layer to fine-tune the text features towards our image-text matching task and generate a text feature embedding $\mathbf{T}_f$ that has the same dimension as the image feature embedding $\mathbf{I}_f$ output by the image encoder. In sum, $\mathbf{T}_f = \Phi^{ITM-R}(\mathbf{T}) = \sigma(\mathbf{W}_R \times \text{Sent-Bert}(\mathbf{T}))$, where $\sigma(\cdot)$ is the sigmoid function.

Loss function Based on the image feature embedding $\mathbf{I}_f$ and the text feature embedding $\mathbf{T}_f$, we minimize a triplet loss similar to that in (Faghri *et al.*, 2018; Lee *et al.*, 2018) as follows.

$$\begin{aligned} \mathcal{L}_{match} = & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)]_+ + \\ & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\bar{\mathbf{I}}_f, \mathbf{T}_f)]_+, \end{aligned} \quad (3.2)$$

where $S(\mathbf{I}_f, \mathbf{T}_f) = \frac{\mathbf{I}_f \cdot \mathbf{T}_f}{\|\mathbf{I}_f\| \cdot \|\mathbf{T}_f\|}$, measuring the cosine similarity between the image and the text features. The symbol $[\cdot]_+$ denotes the function $[x]_+ = \max(x, 0)$; and α serves as a margin parameter. $\bar{\mathbf{T}}_f$ and $\bar{\mathbf{I}}_f$ represent the hardest negative samples (Faghri *et al.*, 2018) in a mini-batch for the positive pair $(\mathbf{I}_f, \mathbf{T}_f)$.

Minimizing this triple loss requires that the distance from $\mathbf{I}_f$ to $\mathbf{T}_f$ be smaller than that from $\mathbf{I}_f$ to $\bar{\mathbf{T}}_f$ by a margin α , and vice versa.

Moreover, it is noticed that negative mining (Faghri *et al.*, 2018) has a crucial impact on the performance of the image-text matching. In our model, we utilize the generated reports from RG-branch as hard negative samples to further improve ITM-branch. As the quality of the generated reports improves over iterations, they become more and more similar to the ground-truth ones, and therefore harder and harder to be differentiated by ITM-branch. When they are used to gradually join the training of image-text matching task, ITM-branch will be enforced to enhance its feature learning so that even finer mismatch between the image and the report could be identified. For this purpose, we additionally introduce a self-boosted triplet loss as follows:

$$\begin{aligned} \mathcal{L}_{match-gen} = & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \mathbf{T}_f^g)]_+ + \\ & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f^g) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)], \end{aligned} \quad (3.3)$$

where $\mathbf{T}_f^g$ refers to the text embedding of the generated report from RG-branch and other symbols remain the same as used in Eqn. 3.2. Minimizing this self-boosted triplet loss enforces ITM-branch to learn effective feature embedding so that in the embedding space, the generated report $\mathbf{T}_f^g$ stays closer to its corresponding image $\mathbf{I}_f$ than other reports $\bar{\mathbf{T}}_f$; at the same time $\mathbf{T}_f^g$ is still farther than its ground-truth $\mathbf{T}_f$ to the image $\mathbf{I}_f$ in order to enforce ITM-branch to differentiate the generated and the ground-truth reports. During our training, the

$\mathcal{L}_{match-gen}$ is only employed after k number of epochs since the generated reports from RG-branch are of poor quality at the early iterations and they cannot assist as hard samples.

3.2.4 Self-boosted Training

The overall objective function of our model combines the loss terms from both RG-branch and ITM-branch:

$$\mathcal{L}_{all} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{ITM}\mathcal{L}_{ITM}, \quad (3.4)$$

where

$$\mathcal{L}_{ITM} = \begin{cases} \mathcal{L}_{match}, & epoch \in (0, k] \\ \mathcal{L}_{match-gen}, & epoch \in (k, N] \end{cases} \quad (3.5)$$

The hyper-parameters λ_{CE} , λ_{feat} , and λ_{ITM} are simply set to balance the three loss terms, and their values are given in Section 3.3. For a detailed experimental analysis of the hyperparameter k , please refer to Table 4.8

To simplify the optimization, we update RG-branch and ITM-branch in an alternative way. That is, in some iterations we fix RG-branch and learn ITM-branch, and in next iterations we fix ITM-branch and learn RG-branch. This pattern is repeated throughout the whole training procedure. Please note that, as the two branches are highly entangled in our model, the update of ITM-branch will affect the visual encoder and the loss of RG-branch; while the update of RG-branch will affect the image encoder of ITM-branch and generate harder samples to the latter. In this way, our training ends up with a self-boosting manner: the improvement of ITM-branch enhances RG-branch and the improvement of RG-branch in turn pushes ITM-branch to evolve.

At the first glance, the interaction between our RG-branch and ITM-branch resembles that observed in Generative Adversarial Networks (GANs). Our RG-branch is like the generator that produces realistic reports while our ITM-branch is like the discriminator that differentiates the generated and the ground-truth reports. However, we would like to point out that our self-boosting mechanism is essentially different from the adversarial training in GANs. First, our RG-branch and ITM-branch do not compete but help with each other. On the contrary, GANs play the two-player min-max game between the generator and the discriminator so that the generator minimizes the reward of the discriminator. Such adversarial loss is not used in our approach. Second, in most GANs (if not all), the generator and the discriminator are two separated networks linked by the adversarial loss, while our RG-branch and ITM-branch share their visual/image encoder as one way of communication.

3.2.5 Remarks

It is noticed that in a very recent work (Zhou *et al.*, 2020), the image-text matching was also used to help image captioning, however, in a substantially different way as ours. In (Zhou *et al.*, 2020) the attentions pretrained by image-text matching are used to regularize the attentions used in report generation, while we communicate the visual and text features between the two tasks throughout the training. Moreover, in (Zhou *et al.*, 2020) the image-text matching and report generation are two separate steps and they are not tightly coupled and interact with each other, which is in a sharp contrast to our self-boosting framework. We also experimentally compare these two methods in Section 3.3. Our results clearly exhibit the superiority of our learning strategy over that of (Zhou *et al.*, 2020) by communicating both the features and the generated samples between the two tasks for progressive mutual improvement.

3.3 Experiments

3.3.1 Datasets

In this experiment, two datasets are used for the performance evaluation. One (IU-Xray) represents the most widely used benchmark for medical report generation, while the other (COV-CTR) is employed to validate model performance on the newly discovered disease COVID-19.

IU-Xray Indiana University Chest X-ray Collection (IU-Xray) (Demner-Fushman *et al.*, 2016) is the most widely used publicly accessible dataset in medical report generation task. It contains 3955 fully de-identified radiology reports, each of which is associated with a frontal and/or lateral chest X-ray images, and 7470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings and Indication etc. In this work, we filtered out reports without two complete image views or without complete sections of “findings” and “impression”, resulting in a smaller dataset with 3195 reports associated with 6390 images. We tokenized all the words in “findings” and “impression” and obtained 2076 unique words. In addition, two special tokens, $\langle start \rangle$ and $\langle end \rangle$, are added to indicate the start and the end of a sentence. We randomly picked 311 (10%) reports to form the test set, which is comparable to the works in the literature (Jing *et al.*, 2018; Xue *et al.*, 2018). All evaluations are done on the test set.

COV-CTR COVID-19 CT Report dataset (COV-CTR)(Li *et al.*, 2023b) contains lungs CT images paired with their corresponding diagnostic chinese reports, in which the lungs CT images are collected by (Yang *et al.*, 2020a) during the outbreak time of COVID-19 and Li et al.(Li *et al.*, 2023b) provided the corresponding diagnostic report, constructing a new medical report dataset. The COV-CTR dataset includes 728 images in total, in which 349 for COVID-19 and

379 for Non-COVID. We used the “jieba” software tool¹ to tokenize all report words, leading to 328 unique words or phrases in total. For a fair comparison, we follow (Li *et al.*, 2023b) to randomly split the data into training, validation, and test sets with the ratio 7:1:2.

3.3.2 Experimental Settings

Evaluation Metrics To evaluate the quality of the generated text report, we utilize the widely used BLEU scores (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004) and CIDER (Vedantam *et al.*, 2015) as evaluation metric. We compute those metrics by the standard image captioning evaluation tool².

Implementation Details We use the paired data as the input for IU-Xray dataset and concatenate the features of frontal and lateral images. The margin parameter α in Eqn. 3.2 and 3.3 is set to 0.2 and $\lambda_{CE}, \lambda_{feat}, \lambda_{ITM} = 2, 10, 1$ in Eqn. 3.4, respectively. The word embedding size and topic vector size were respectively set to 300 and 512. We train our model using Adam optimizer (Kingma and Ba, 2015) with mini-batch size of 16. The learning rate are set to be 0.0001 and 0.0002, respectively, for RG-branch and ITM-branch with a total 30 epochs. We set $k = 10$ in Eqn. 3.3, training ITM branch with the regular triplet loss for the first 10 epochs and the self-boosted triple loss for the rest.

3.3.3 Results and Discussion

Comparison with SOTA We compare our model with 7 state-of-the-art image captioning methods. These include the classic model Show-tell (Vinyals *et al.*, 2015), different attention-based methods including AdaAtt (Lu *et al.*, 2017), Att2in (Anderson *et al.*, 2018), and Topdown (Rennie *et al.*, 2017), methods using advanced NLP models including Transformer (Devlin *et al.*,

¹<https://github.com/fxsjy/jieba>

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

Table 3.1. Comparison on IU-Xray (upper part) and COV-CTR datasets (lower part). † indicates the results are quoted from the published literature. Specifically, we quote the results from (Li *et al.*, 2019a) for IU-xray, and from (Li *et al.*, 2023b) for COV-CTR. For other methods in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
IU-Xray	Show-Tell	0.346	0.214	0.141	0.095	0.320	0.239
	Att2in	0.399	0.249	0.172	0.126	0.321	0.341
	AdaAtt	0.436	0.288	0.203	0.150	0.354	0.265
	Topdown	0.375	0.237	0.167	0.123	0.319	0.336
	Transformer	0.422	0.264	0.177	0.120	0.338	0.268
	M2transformer	0.463	0.318	0.214	0.155	0.335	0.349
	Grounded	0.446	0.301	0.237	0.176	0.343	0.395
	CoAtt†	0.455	0.288	0.205	0.154	0.369	0.277
	HGRG-Agent†	0.438	0.298	0.208	0.151	0.322	0.343
	KERP†	0.482	0.325	0.226	0.162	0.339	0.280
	Ours	0.487	0.346	0.270	0.208	0.359	0.452
COV-CTR	Show-Tell	0.665	0.613	0.578	0.549	0.684	0.631
	Att2in	0.674	0.614	0.574	0.542	0.588	0.686
	AdaAtt	0.689	0.636	0.596	0.563	0.709	0.729
	Topdown	0.705	0.663	0.622	0.586	0.720	0.968
	Transformer	0.683	0.636	0.590	0.558	0.719	0.721
	M2transformer	0.733	0.662	0.620	0.582	0.750	1.289
	Grounded	0.753	0.708	0.665	0.627	0.776	1.381
	CoAtt†	0.709	0.645	0.603	0.552	0.718	0.672
	Vision-BERT†	0.710	0.653	0.606	0.558	0.747	0.684
	ASGK†	0.712	0.659	0.611	0.570	0.746	0.680
	Ours	0.810	0.766	0.721	0.679	0.790	2.371

2019) and M2transformer (Cornia *et al.*, 2020), and the very recent method Grounded (Zhou *et al.*, 2020) that also takes advantage of the image-text matching task to improve image captioning but in a completely different manner as ours. For these methods, we download the codes released publicly and re-run them on these two datasets with the same experiment setting as ours. Moreover, we also compare with 3 state-of-the-art medical report generation models: CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et al.*, 2018), KERP (Li *et al.*, 2019a). Since medical report generation models do not release the codes, we have to quote their performance from (Li *et al.*, 2019a) that is often referred to in this field.

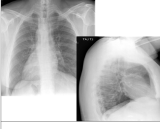
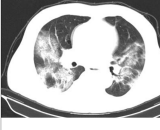
Input Images	Ground Truth	Show-Tell	Grounded	Ours
	Impression: No acute cardiopulmonary finding. Findings: The heart size and cardio mediastinal silhouette are normal. There is no focal air space opacity, pleural effusion, or pneumothorax. The osseous structures are intact with mild degenerative changes in thoracic spine	Impression: no acute cardiopulmonary abnormality . Findings: lungs are clear bilaterally . cardiac and mediastinal silhouettes are normal . pulmonary vasculature is normal . no pneumothorax or pleural effusion . no acute bony abnormality	Impression: no acute cardiopulmonary abnormality . Findings: the lungs are clear and without focal air space opacity . cardio mediastinal silhouette is normal in size and contour and stable . there is no pneumothorax or large pleural effusion	Impression: no acute cardiopulmonary finding . Findings: the cardio mediastinal silhouette is normal in size and contour . the lungs are clear . there is no pleural effusion or pneumothorax . the osseous structures are intact with mild degenerative changes in thoracic spine
	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph nodes are found in mediastinum, bilateral lung markings are enhanced, patchy ground glass shadows are seen in both lungs, the boundary is fuzzy, some density is uneven, lobar bronchus is unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph node shadow is found in mediastinum, no abnormal density shadow is found in both lungs, lobar bronchus is unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity.	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph nodes are found in mediastinum, bilateral lung markings are enhanced, patchy shadows are seen in right lower lobe with clear boundary, lobar bronchus are unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity.	The chest is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph node shadow is found in mediastinum, bilateral lung markings are enhanced, multiple flake ground glass shadows are seen in both lungs, the boundary is fuzzy, the bilateral hilar is not big, some of the density is uneven, the lobar bronchus are unobstructed, and no abnormal density shadow is found in bilateral chest.

Figure 3.2. Examples of the generated report on IU-xray(upper part) and COV-CTR dataset (lower part). We compare our results with the ground truth report and the other two methods: Show-Tell (Vinyals *et al.*, 2015) and Grounded (Zhou *et al.*, 2020). Highlighted sentences are the key information in the ground truth report that is only generated by our method.

As shown in Table 3.1, on both datasets, our self-boosting method achieves the best performance over almost all evaluation metrics among the comparing methods. The two very recent image captioning methods M2transformer (Cornia *et al.*, 2020) and Grounded (Zhou *et al.*, 2020) perform better than other existing methods, but still lose to ours. Especially, our advantages over Grounded (Zhou *et al.*, 2020) fully demonstrate the benefit to jointly train the two tasks of report generation and image-text matching in our proposed self-boosting manner. It is also noted that, our performance improvement on CIDEr is most significant. Cross-referring to Table 3.2, even without the reinforcement learning on CIDEr (denoted as Ours-RL), our method is still the best performer on CIDEr on both datasets and the advantage is pronounced³. This is a positive sign for radiographic reports, since CIDEr focuses on key information by down-weighting the common words in all reports.

In addition, we also conduct qualitative comparison among our method, the classic ShowTell model (Vinyals *et al.*, 2015), and the most recent and closest method of Grounded (Zhou *et al.*, 2020). Two examples of the generated reports are visually compared in Figure 3.2, one from each of the two datasets. A

³It is noted that Grounded (Zhou *et al.*, 2020) also uses RL on CIDEr for refinement.

wider visual comparison among more examples and methods is provided in our supplement. Please note that COV-CTR reports are generated and evaluated in Chinese. They are simply translated into English by Google Translate⁴ for display. As can be observed, our proposed model can significantly improve the quality of the generated reports, consistent with the above quantitative analysis. For example, our generated report correctly describes “the osseous structures” in the first example, and points out “the boundary is fuzzy” and “the density is uneven” in the second example, while the other two methods missed such information.

Dataset	Model	B@4	Rouge	CIDEr
IU-Xray	baseline	0.128	0.307	0.359
	baseline+ITM	0.155	0.321	0.372
	baseline+ITM+FLL	0.169	0.330	0.391
	baseline+ITM+STL	0.176	0.345	0.418
	Ours-RL	0.193	0.352	0.426
	Ours	0.208	0.359	0.452
COV-CTR	baseline	0.573	0.659	0.722
	baseline+ITM	0.611	0.704	1.243
	baseline+ITM+FLL	0.637	0.729	1.617
	baseline+ITM+STL	0.658	0.761	1.953
	Ours-RL	0.672	0.783	2.014
	Ours	0.679	0.790	2.371

Table 3.2. Ablation studies. “Baseline” refers to RG-branch only. “ITM”, “FLL”, “STL” and “RL” stand for “Image-Text Matching”, “Feature Level Loss” ($\mathcal{L}_{feat}$ in Eqn. 3.1), “Self-booster Triplet Loss” (Eqn. 3.3), and “Reinforcement Learning”, respectively. B@4 is the BLEU score uses 4-grams.

Ablation Study We conduct an ablation study to single out the contributions of each component in our proposed method, as presented in Table 3.2. We take the RG branch of our proposed pipeline as the baseline to verify the performance improvements brought by ITM-branch and our self-booster training strategy. In Table 3.2, there are four components: ITM, FLL, STL and RL, representing Image-Text matching, Feature-Level Loss, Self-booster Loss and Reinforcement Learning, respectively. The symbols “+” or “-” indicate the inclusion or exclusion of the following component. The benefit of using the auxiliary

⁴<https://translate.google.com/>

image-text matching task can be well reflected by the improvement from “baseline” to “baseline+ITM”. As shown, the performance can be further boosted by additionally utilizing feature-level loss (“baseline+ITM+FLL”) in RG-branch and the self-boosted triplet loss (“baseline+ITM+STL”) in ITM-branch, proving their indispensability in our proposed method. Moreover, as mentioned, even removing the refinement by reinforcement learning, our model still significantly outperforms the comparing methods in Table 3.1.

Hyper-Parameter Study A sensitivity analysis was conducted to assess the impact of hyperparameters within the IU-Xray dataset. We fixed $\lambda_{ITM} = 1$ and adjusted λ_{CE} and λ_{feat} by a factor of 1.5, both increasing and decreasing. According to the results presented in Table 3.3, our method consistently outperforms all state-of-the-art (SOTA) methods listed in Table 3.1 on the IU-Xray dataset, even when these hyperparameters are varied within a significantly broad range.

Table 3.3. Hyper-parameter study using IU-Xray ($\lambda_{ITM} = 1$) . B@4 is the BLEU score uses 4-grams.

λ_{CE}	λ_{feat}	B@4	Rouge	CIDEr
2.0	10.0	0.208	0.359	0.452
1.3 (2 / 1.5)	10.0	0.194	0.351	0.416
3.0 (2×1.5)	10.0	0.197	0.349	0.412
2.0	6.7 (10 / 1.5)	0.195	0.348	0.413
2.0	15.0 (10×1.5)	0.203	0.352	0.421
Best SOTA (Grounded (Zhou <i>et al.</i> , 2020))		0.176	0.343	0.395

3.4 Conclusions

In this work, we improve the diagnostic report generation by additionally utilizing an auxiliary image-text matching task to learn strongly correlated visual and text features to describe fine-grained differences among radiographic images. We show how to achieve this by deeply coupling the two tasks and encouraging the cooperative interactions between them. By communicating the learned features and the newly generated samples, the two tasks could mutually boost each

other in a progressive way. As experimentally verified, our generated reports could better capture the subtle but critical information in radiographic images.

Diagnostic Captioning by Cooperative Task Interactions and Sample-graph Consistency

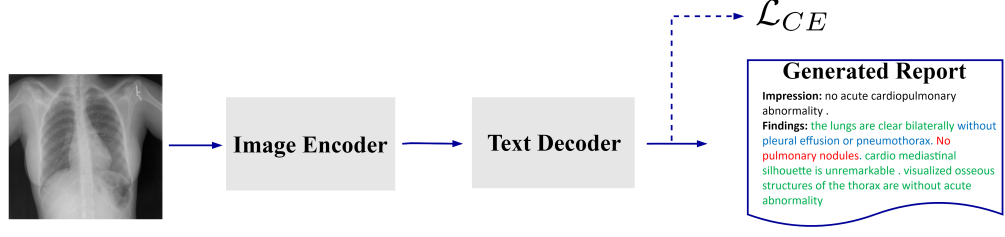
Radiographic images are similar to each other, making it challenging for diagnostic captioning to narrate fine-grained visual differences of clinical importance. In this chapter, we propose a self-boosting framework integrating two novel strategies to learn tightly correlated image and text features for diagnostic captioning. The first strategy explicitly aligns image and text features through training an auxiliary task of image-text matching (ITM) jointly with the main task of report generation (RG) as two branches of a network model. The ITM branch explicitly learns image-text alignment and provides highly correlated visual and textual features for the RG branch to generate high-quality reports. The high-quality reports generated by RG branch, in turn, are utilized as additional harder negative samples to push the ITM branch to evolve towards better image-text alignment. These two branches help improve each other progressively, so that the whole model is self-boosted without requiring external resources. The second strategy aligns image-sample space and report-sample space to achieve consistent image and text feature embeddings. To achieve this, the sample graph of the embedded ground-truth reports is built and used as the target to train the sample graph of the embedded images so that the fine discrepancy in the ground-truth reports could be captured by the learned visual feature embeddings. Our proposed framework demonstrates its superiority on two medical report generation benchmarks, including the largest dataset MIMIC-CXR.

4.1 Introduction

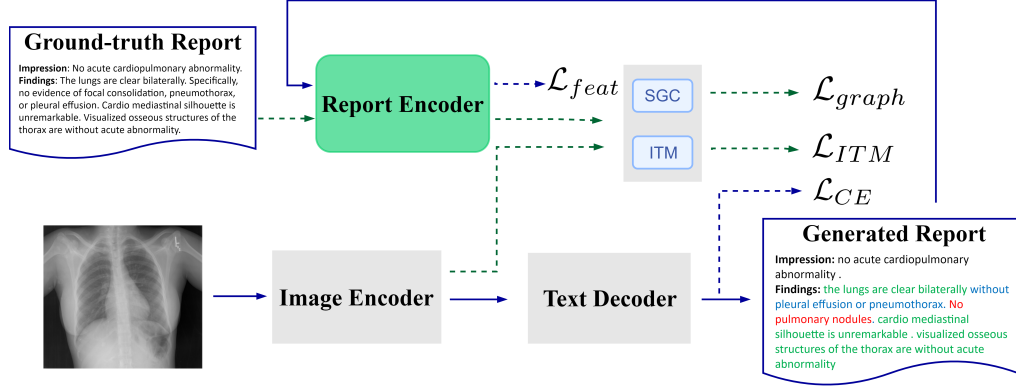
Building upon the foundational work outlined in Chapter 3, this chapter aims to further evolve our self-boosting framework for automated radiographic report generation. While Chapter 3 laid the groundwork by introducing the auxiliary task of Image-Text Matching (ITM) in conjunction with the main task of Report Generation (RG), this chapter introduces innovative extensions and optimizations to tackle the challenges associated with fine-grained diagnostic captioning.

In this context, several advancements stand out. First, we replace the hierarchical LSTM used as the text decoder in the Chapter 3 with a more powerful memory-augmented Transformer. This modification is geared to make our framework more adept at nuanced text generation, aligning more closely with the complexity of clinical language. Second, in response to some of the limitations observed in Chapter 3, we introduce a novel sample-graph consistency loss. This element of the model is inspired by the benefits observed in retrieval-based approaches in the broader image captioning domain. This sample-graph consistency loss serves as an additional regularization layer within our self-boosting framework and is designed to further fine-tune the alignment between visual and text features. In addition, we extend our evaluation metrics to include MIMIC-CXR, the largest newly published chest X-ray dataset. This allows us to scrutinize our model’s efficacy under more challenging conditions and against more competitive baselines. We also delve into new ablation studies, examining the impact of various hyper-parameters that were not addressed in the previous chapter.

Comparing with Chapter 3, the main contributions of this chapter are summarized as follows.



(a) common medical report generation approach



(b) Proposed self-boosting framework

Figure 4.1. Compared with the common encoder-decoder structure used for report generation (a), our self-boosting framework for diagnostic captioning (b) jointly trains two tasks: the main task of report generation (RG), and the auxiliary task of image-text matching (ITM), built as two branches of a deep learning model. The RG branch (indicated by the solid line) takes an image as the input and outputs the generated report. The ITM branch (indicated by the dash line) takes an image-report pair as the input and predicts whether they are a true match. The RG-branch and the ITM branch share the same image encoder. They gradually help improve each other through their cooperative interactions.

1) Building upon the hierarchical LSTM text decoder introduced in Chapter 3, this chapter explores two additional backbones. One employs the memory-augmented Transformer as the text decoder to enhance medical report generation, while the other incorporates our proposed strategies into the X-Transformer backbone to assess their efficacy in general image captioning tasks. Both show marked improvements in our experimental evaluations.

2) Extending the self-boosting strategy initially described in Chapter 3, this chapter introduces a novel sample-graph consistency loss. This loss function

leverages sample relationships to regularize interactions between image and report features. Its effectiveness has been validated through empirical studies and has been integrated into our self-boosting framework in this chapter.

3) We expand our experimental study to include the largest chest X-ray dataset MIMIC-CXR (Johnson *et al.*, 2019), compare our method with more competitive approaches, and assess the clinical correctness of the generated reports using the medical terms in Radgraph, which were not considered in Chapter 3.

4) We conduct new ablation studies to investigate the impact of hyper-parameters to our model performance, including those parameters balancing loss terms, the parameter controlling the alternating optimisation, and the parameter switching the two triplet losses. We also add visualisation to show the effectiveness of the newly proposed sample-graph consistency loss.

5) Last but not least, we also validate our proposed two strategies for general image captioning task in the Chapter. Specifically, we incorporate our strategies to the X-Transformer backbone, and evaluate its performance on a large image captioning benchmark COCO. The experimental results demonstrate that our model outperforms many state-of-the-arts image captioning models as well.

4.2 Method

In this section, we first give an overview of the proposed model. Then two important branches of our model, i.e., image-text matching and medical report generation, are presented. Following that, we describe how these two branches achieve joint training and mutual improvement. At last, to further augment our self-boosting framework, we introduce a sample-graph consistency loss and describe how this loss can drive the feature encoders of the two modalities to learn consistent visual and text embeddings.

4.2.1 Overview

A typical radiographic report consists of a conclusive “impression” part and a descriptive “finding” part. Taking a medical image as input, our model generates a long-paragraph diagnostic report with both parts. As mentioned in Section 4.1, our model improves report generation by taking advantage of the interactions between the main task of report generation and the auxiliary task of image-text matching, while the report generation task is additionally supervised by a sample-graph consistency loss to ensure consistent visual and text feature embeddings. Our model is trained in a self-boosting manner.

Figure 4.2 provides an overview about our framework. In general, our model consists of two entangled branches: medical report generation (main task) and image-text matching (auxiliary task). As for the main task, our report generation branch (RG-branch) follows the encoder-decoder architecture. The visual encoder detects image regions with an unsupervised method, extracts regional features using CNN, and refines these visual features based on the regional relationships learnt through self-attention. The text decoder takes the refined visual features as input, passes them through a memory-augmented Transformer decoder or a hierarchical LSTM model catering for both the topic-level and word-level decoding, and generates the final paragraphical report. As for the auxiliary task, taking the image-report pairs as the input, our image-text matching branch (ITM-branch) consists of an image encoder and a report encoder, learning the visual and the text features, respectively, in order to predict if the input image and report match each other.

The two branches interact with each other in three ways. First, they share the visual encoder, so that the text-correlated visual features learned by ITM-branch could also be utilized to improve the RG-branch to generate high-quality reports. Second, the report encoder learnt by the ITM-branch is also utilized by the RG-branch to extract textual features to i) evaluate and minimize the feature-level

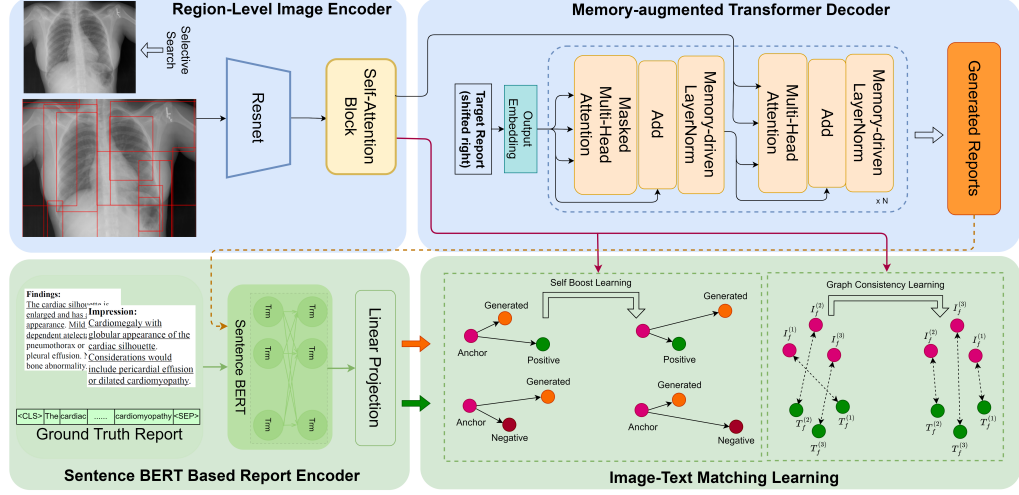


Figure 4.2. An overview of the network structure of the proposed framework with a Transformer-based decoder. It comprises of a medical report generation branch (RG-branch, the upper part) and an auxiliary image-text matching branch (ITM-branch, the lower part). RG-branch consists of two components: a region-level image encoder (shared with ITM-branch) to extract visual features from medical images, and a memory-augmented Transformer decoder to generate diagnose reports. ITM-branch consists of an image encoder (shared with RG-branch) and a report encoder based on Sentence-BERT. It compares the similarity of visual and text features for match. The two branches are designed to have cooperative interactions during training. ITM-branch provides RG-branch text-correlated visual features for report generation and its report encoder for measuring the feature level loss of the generated reports. In turn, with the improved generated reports, RG-branch supplies ITM-branch harder and harder samples to force the latter to evolve for better feature learning. In this way, the whole model is self boosted.

loss between the generated and the ground-truth reports, contributing to our feature-level loss $\mathcal{L}_{feat}$; and ii) regularize the consistency of the visual and the textual feature embeddings, contributing to our sample graph consistency loss $\mathcal{L}_{graph}$. Third, during the training, the improved reports generated by RG-branch are passed to the ITM-branch as additional harder samples, which enforces the ITM-branch to improve itself by enhancing the feature learning. These three interactions last for the training course, making the whole model gradually boost itself. At inference stage, only RG-branch is used, which receives a test image and generates the corresponding report.

4.2.2 Report Generation (RG) Branch

The ultimate target of our model is to generate ground-truth-like diagnostic reports from radiographic images. Therefore, report generation is our main task. Like most image-captioning methods, our RG-branch consists of a visual encoder and a successive text decoder, with enhanced characteristics. They are elaborated as follows.

Visual encoder Although some generic image captioners use Faster R-CNN (Ren *et al.*, 2017a) to encode image regions at object-level, existing medical report generation methods extract visual features from the whole image rather than image regions, due to the lack of supervised information. Differently, given an input image of size $3 \times H \times W$, we employ the selective search algorithm to generate region proposals unsupervisely, and then refine these region proposals by: 1) excluding regions smaller than 2000 pixels; 2) excluding regions where the background ratio exceeds 70%. This leads to roughly $M = 30$ regions for one image. After that, we adopt ResNet-101 to produce a set of vectors $v_1, \dots, v_M \in \mathbb{R}^D$ with a dimension $D = 2048$ for each region. Compared with image-level features, using regional features can “look closer” to the image, thus better suits the fine-grained pattern description. It is found that modeling region relationship is beneficial for visual representation. Unlike some image-captioning methods (Yao *et al.*, 2018; Yang *et al.*, 2019) that rely on the Visual Genome dataset to learn GCN or scene graph for generic images, we employ self-attention (Vaswani *et al.*, 2017) to learn the relationships among regions and transform regional visual features $\mathbf{V}$ into relationship-enhanced ones $\mathbf{V}_h$ for refinement. These relationship-enhanced regional features are then passed through a region pooling layer to generate the pooled vector $\mathbf{v}_p$. We further input the pooled vector $\mathbf{v}_p$ through a 2048-dimensional fully-connected layer to produce the final representation: $\mathbf{I}_f = \mathbf{W}_f \mathbf{v}_p + \mathbf{b}_f$. The output of the visual encoder $\mathbf{I}_f$ is then sent to the text decoder in RG-branch for report generation, as well as to ITM-branch for predicting image-text matching.

Text decoder To generate paragraphical report, we explore two different structures decoder model: a hierarchical structured LSTM (HLSTM) and a memory augmented transformer (MATM) (Chen *et al.*, 2020b). HLSTM includes a sentence LSTM and a word LSTM (Krause *et al.*, 2017). The sentence LSTM takes the visual features as input, produces a *topic vector* for each sentence in the report paragraph, and determines when to stop generating the topic vector. Those topic vectors are then fed into the word LSTM to produce the sentences. MATM is built by integrating a memory-bank module into Transformer which can better record the information from previous processes to cater for long paragraph generation. More details are as follows.

HLSTM. For the hierarchical LSTM, the sentence LSTM is a single layer LSTM with a hidden size $H = 512$. It takes the image feature $\mathbf{I}_f$ as input, and in turn generates a sequence of hidden states $\mathbf{h}_1 \cdots \mathbf{h}_N \in \mathbb{R}^H$, one for each sentence in the paragraph: $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{I}_f, (\mathbf{h}_i, \mathbf{c}_i))$, $i \in \{0, \dots, N-1\}$, where $\mathbf{c}_i$ denotes the memory cells of LSTM, and N is the number of sentences in the paragraph. Based on the hidden state $\mathbf{h}_i$, a topic vector $\mathbf{t}_i \in \mathbb{R}^P$ is obtained as: $\mathbf{t}_i = \tanh(\mathbf{W}_t \mathbf{h}_i + \mathbf{b}_t)$ $i \in \{1, \dots, N\}$. In addition, the hidden state $\mathbf{h}_i$ is also mapped from H dimension to two dimension by a linear projection to produce a probability q_i through a sigmoid function, determining whether the i th sentence is the last sentence in the paragraph. q_i can be learnt by minimizing the cross entropy loss for a binary classification of being continued or not.

The word LSTM is also a single layer LSTM with hidden size $H = 512$ and the initial hidden and cell states of the word LSTM are set to zero. It takes the topic vectors produced by the sentence LSTM as the input and generates detailed sentences. Given the i th topic $\mathbf{t}_i$, the hidden and cell states of the word LSTM are updated by $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{s}_{j-1}^{(i)}, (\mathbf{h}_i, \mathbf{c}_i))$, where $j \in \{0, \dots, M+2\}$, and M is the number of word in i th sentence. It should be noted, when $j = 0$, $\mathbf{s}_{-1}^{(i)}$ represents the topic vector $\mathbf{t}_i$; when $j > 0$, $\mathbf{s}_j^{(i)}$ represents the word embedding of the j th word in the i th sentence. In particular, $\mathbf{s}_0$ and $\mathbf{s}_{M+1}$ represent the word

embedding of the special token $\langle start \rangle$ and $\langle end \rangle$ respectively. After obtaining the hidden state of word LSTM, the probability $p_{ij} = \text{Softmax}(\mathbf{W}_p \mathbf{h}_j + \mathbf{b}_p)$ can be learnt, predicting the probability of j th word in i th sentence, where $i \in \{1, \dots, M\}, j \in \{1, \dots, N\}$.

MATM. For the memory-augmented Transformer decoder, it is built upon the vanilla Transformer (Vaswani *et al.*, 2017), including two critical components: relational memory module (RM) and memory-driven conditional layer normalization module (MCLN). The RM module is designed to explore the relation between image-report pairs and record the information computed during the generation process, while the MCLN module is designed to integrate the RM module into Transformer. Mathematically, the t -th word in generated report can be obtained by: $y_t = \text{MAMT}(\mathbf{V}_h, \text{MCLN}(\text{RM}(y_1, y_2, \dots, y_{t-1})))$, where MAMT refers to the memory-augmented Transformer decoder, $\mathbf{V}_h$ is the relationship-enhanced visual features extracted from the visual encoder, and y_i represents the i th predicted word. More details about RM and MCLN module could be referred to the work in (Chen *et al.*, 2020b).

Loss function Most image captioning and medical report generation methods construct the loss functions evaluating the difference between the generated and the ground-truth reports at the word-level (cross-entropy based loss). In this chapter, we further argue that a generated report with high-quality should also be similar to the ground-truth at the semantic feature levels. That is, they should stay similar after they are passed through a text feature extractor producing features corresponding to high-level semantics. In our work, such a text feature extractor could be readily achieved by utilizing the report encoder learnt in the auxiliary ITM-branch, which is well trained to effectively encode the most critical text features correlated to the visual contents. Specifically, given K reports, the loss function used in our RG-branch consists of three terms as follows:

$$\mathcal{L}_{RG} = \lambda_{CE} \sum_{k=1}^K \mathcal{L}_{CE}^{(k)} + \lambda_{feat} \sum_{k=1}^K \mathcal{L}_{feat}^{(k)} + \lambda_{gc} \sum_{k=1}^K \mathcal{L}_{graph}^{(k)}, \quad (4.1)$$

where λ_{CE} , λ_{feat} and λ_{graph} are hyper-parameters balancing the three terms.

The first loss term $\mathcal{L}_{CE}^{(k)}$ is the cross-entropy loss used to measure the word-level difference for the k th report. Since we use a hierarchical LSTM as the text decoder, the $\mathcal{L}_{CE}^{(k)}$ is the sum of the cross-entropy loss $\mathcal{L}_S^{(k)}$ of the sentence LSTM and the cross-entropy loss $\mathcal{L}_W^{(k)}$ of the word LSTM. The loss $\mathcal{L}_S^{(k)} = -\sum_{i=0}^N \left[l_i^{(k)} \cdot \log(q_i^{(k)}) + (1 - l_i^{(k)}) \cdot \log(1 - q_i^{(k)}) \right]$, where $q_i^{(k)}$ and $l_i^{(k)}$ are the predicted and the ground-truth labels, respectively, indicating if the i th sentence is the last sentence in the k th report. The loss $\mathcal{L}_W^{(k)} = -\sum_{i=1}^N \sum_{j=1}^M y_{ij}^{(k)} \log(p_{ij}^{(k)})$, where $y_{ij}^{(k)}$ refers to the ground-truth and $p_{ij}^{(k)}$ refers to the prediction of the j th word in the i th sentence within the k th report.

The second loss term $\mathcal{L}_{feat}^{(k)} = \|\Phi^{ITM-R}(\hat{\mathbf{T}}^{(k)}) - \Phi^{ITM-R}(\mathbf{T}^{(k)})\|_2$ measures the difference at the feature level, where $\hat{\mathbf{T}}^{(k)}$ and $\mathbf{T}^{(k)}$ are the predicted paragraph and the ground-truth for the k th report, respectively. Here the mapping $\Phi^{ITM-R}(\cdot)$ denotes the embedding function of the report encoder in ITM-branch (see Section 4.2.3 for more details), which is highly nonlinear.

The third loss term $\mathcal{L}_{graph}^{(k)}$ measures the consistency between the image sample-graph and the (ground-truth) report sample-graph constructed in their respective feature embedding spaces. Please see Section 4.2.5 for details.

In addition, following literature (Rennie *et al.*, 2017), we also apply reinforcement learning with CIDEr as the reward to further boost the performance of our RG-branch.

4.2.3 Image-Text Matching (ITM) Branch

To promote the main task of report generation, we also introduce an auxiliary task of image-text matching to learn highly correlated visual and text features. Our ITM-branch takes an image-report pair as input and tells whether the image

and the report match. It consists of an image encoder and a report encoder, detailed as follows.

Image encoder Our ITM-branch shares its image encoder with our RG-branch. As mentioned above, the image encoder extracts visual features from regions, refines these features with regional relationship, and outputs the image representation on top of an additional region pooling. It is worth mentioning that, sharing the image encoder between ITM-branch and RG-branch enables our model to both learn image-text correlated features and fine-tune these features for the main report generation task.

Report encoder We considered BERT’s (Devlin *et al.*, 2019) model as the report encoder, which is a multi-layer bidirectional Transformer (Vaswani *et al.*, 2017) and has set a new state-of-the-art performance on various NLP tasks. However, we do not directly use the vanilla BERT. As mentioned, the report encoder in ITM-branch is also employed by RG-branch to compare the feature similarity between the generated and the ground-truth reports. Taking this into account, we adapt a Sentence-BERT (Reimers and Gurevych, 2019) model to constructing our report encoder, because Sentence-BERT employed a siamese and triplet network structure (Schroff *et al.*, 2015) for BERT, making it more suitable to derive semantically meaningful sentence embeddings. Sentence-BERT has been well pre-trained on two large broad-coverage corpus (Bowman *et al.*, 2015; Williams *et al.*, 2018) for various textual similarity tasks. Moreover, on top of the pretrained Sentence-BERT, we stack another fully connected layer to fine-tune the text features towards our image-text matching task and generate a text feature embedding $\mathbf{T}_f$ ¹ that has the same dimension as the image feature embedding $\mathbf{I}_f$ output by the image encoder. In sum, $\mathbf{T}_f = \Phi^{ITM-R}(\mathbf{T}) = \sigma(\mathbf{W}_R \times \text{Sent-Bert}(\mathbf{T}))$, where $\sigma(\cdot)$ is the sigmoid function.

¹The subscript “f” in T_f and I_f indicates the feature space, instead of the sample index.

Loss function Based on the image feature embedding $\mathbf{I}_f$ and the text feature embedding $\mathbf{T}_f$, we minimize a triplet loss similar to that in (Faghri *et al.*, 2018; Lee *et al.*, 2018) as follows.

$$\begin{aligned} \mathcal{L}_{match} = & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)]_+ + \\ & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\bar{\mathbf{I}}_f, \mathbf{T}_f)]_+, \end{aligned} \quad (4.2)$$

where $S(\mathbf{I}_f, \mathbf{T}_f) = \frac{\mathbf{I}_f \cdot \mathbf{T}_f}{\|\mathbf{I}_f\| \cdot \|\mathbf{T}_f\|}$, measuring the cosine similarity between the image and the text features. The symbol $[\cdot]_+$ denotes the function $[x]_+ = \max(x, 0)$; and α serves as a margin parameter. $\bar{\mathbf{T}}_f$ and $\bar{\mathbf{I}}_f$ represent the hardest negative samples (Faghri *et al.*, 2018) in a mini-batch for the positive pair $(\mathbf{I}_f, \mathbf{T}_f)$. Minimizing this triple loss requires that the distance from $\mathbf{I}_f$ to $\mathbf{T}_f$ be smaller than that from $\mathbf{I}_f$ to $\bar{\mathbf{T}}_f$ by a margin α , and vice versa.

Moreover, it is noticed that negative mining (Faghri *et al.*, 2018) has a crucial impact on the performance of the image-text matching. In our model, we utilize the generated reports from RG-branch as hard negative samples to further improve ITM-branch. As the quality of the generated reports improves over iterations, they become more and more similar to the ground-truth ones, and therefore harder and harder to be differentiated by ITM-branch. When they are used to gradually join the training of image-text matching task, ITM-branch will be enforced to enhance its feature learning so that even finer mismatch between the image and the report could be identified. For this purpose, we additionally introduce a self-boosted triplet loss as follows:

$$\begin{aligned} \mathcal{L}_{match-gen} = & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \mathbf{T}_f^g)]_+ + \\ & [\alpha - S(\mathbf{I}_f, \mathbf{T}_f^g) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)], \end{aligned} \quad (4.3)$$

where $\mathbf{T}_f^g$ refers to the text embedding of the generated report from RG-branch and other symbols remain the same as used in Eqn. 4.2. Minimizing this self-boosted triplet loss enforces ITM-branch to learn effective feature embedding

so that in the embedding space, the generated report $\mathbf{T}_f^g$ stays closer to its corresponding image $\mathbf{I}_f$ than other reports $\bar{\mathbf{T}}_f$; at the same time $\mathbf{T}_f^g$ is still farther than its ground-truth $\mathbf{T}_f$ to the image $\mathbf{I}_f$ in order to enforce ITM-branch to differentiate the generated and the ground-truth reports. During our training, the $\mathcal{L}_{match-gen}$ is only employed after k number of epochs since the generated reports from RG-branch are of poor quality at the early iterations and they cannot assist as hard samples.

4.2.4 Self-boosted Training

The overall objective function of our model combines the loss terms from both RG-branch and ITM-branch:

$$\mathcal{L}_{all} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{SGCL}\mathcal{L}_{graph} + \lambda_{ITM}\mathcal{L}_{ITM}, \quad (4.4)$$

where

$$\mathcal{L}_{ITM} = \begin{cases} \mathcal{L}_{match}, & epoch \in (0, k] \\ \mathcal{L}_{match-gen}, & epoch \in (k, N] \end{cases} \quad (4.5)$$

The hyper-parameters λ_{CE} , λ_{feat} , λ_{SGCL} and λ_{ITM} are simply set to balance the four loss terms, and their values are given in Section 4.3.

To simplify the optimization, we update RG-branch and ITM-branch in an alternative way. Specifically, we trained the ITM-branch for a fixed number of epochs, denoted as m , and subsequently trained the RG-branch for the same number of epochs, with this pattern being repeated iteratively throughout the training process. In our experiments, we set the value of m to 5. Please note that, as the two branches are highly entangled in our model, the update of ITM-branch will affect the visual encoder and the loss of RG-branch; while the update of RG-branch will affect the image encoder of ITM-branch and generate harder

samples to the latter. In this way, our training ends up with a self-boosting manner: the improvement of ITM-branch enhances RG-branch and the improvement of RG-branch in turn pushes ITM-branch to evolve.

At the first glance, the interaction between our RG-branch and ITM-branch resembles that observed in Generative Adversarial Networks (GANs). Our RG-branch is like the generator that produces realistic reports while our ITM-branch is like the discriminator that differentiates the generated and the ground-truth reports. However, we would like to point out that our self-boosting mechanism is essentially different from the adversarial training in GANs. First, our RG-branch and ITM-branch do not compete but help with each other. On the contrary, GANs play the two-player min-max game between the generator and the discriminator so that the generator minimizes the reward of the discriminator. Such adversarial loss is not used in our approach. Second, in most GANs (if not all), the generator and the discriminator are two separated networks linked by the adversarial loss, while our RG-branch and ITM-branch share their visual/image encoder as one way of communication.

4.2.5 Sample-graph Consistency Loss

As introduced in Section 4.1, it is difficult for the current image captioning or diagnostic captioning methods to perceive two very similar images with essential difference in their reports. To further improve the perceptual resolution of diagnostic captioning models, we propose a sample-graph consistency loss to align image metric space and report metric space to obtain consistent visual and textual embeddings, so that the distance between two report sample embeddings could be reflected by the distance between their corresponding image sample embeddings. The basic idea is illustrated in Figure 4.3. Specifically, we first build both the image sample-graph (with image sample embeddings as the nodes) and the report sample-graph (with report sample embeddings as the nodes) to reflect sample relationship. By requiring the two graphs to have the

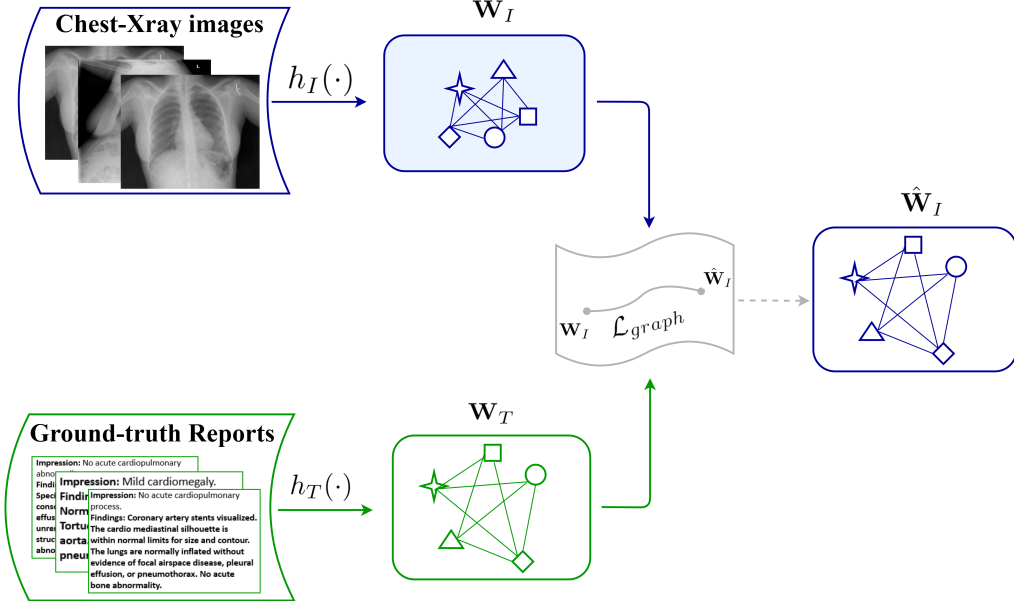


Figure 4.3. Illustration of our sample-graph consistency loss. Blue and green indicate the image modality and the report modality, respectively. The images and the reports are mapped into their respective feature spaces by the image encoder h_I and the report encoder h_T . In the feature spaces, both the sample-graph $\mathbf{W}_I$ of the images and the sample-graph $\mathbf{W}_T$ of the reports are constructed. The graph nodes sharing the same symbol in $\mathbf{W}_I$ and $\mathbf{W}_T$ belong to one image-report pair and the edge length between two nodes (i.e., samples) of a graph represents their distance in the embedding space. Initially, in the image feature space, the embeddings of two Chest-Xray images may be very close due to the fine-grained visual difference, despite the essential difference in their corresponding reports. Our sample-graph consistency loss takes the sample-graph $\mathbf{W}_T$ of the ground-truth reports as the supervision, and minimizes the difference between the two sample-graphs $\mathbf{W}_T$ and $\mathbf{W}_I$ so that two similar images with different ground-truth reports could stay apart in the feature embedding space.

same structure, we utilise the graph of the embedded ground-truth reports as the target to train the graph of the embedded images. In this way, two similar but different ground-truth reports will correspond to two close but different visual embeddings. Also, as the generated reports are regularized by the ground-truth reports in the training set, they could better conform to the distribution of the ground-truth reports. Considering the computational efficiency, the graph-consistency loss is evaluated at batch level. The proposed approach is elaborated as follows.

Let us consider a batch of B training samples $\chi = \{\mathbf{I}^i, \mathbf{T}^i\}_{i=1}^B$, where $\mathbf{I}^i$ and $\mathbf{T}^i$ refer to the image and the report of the i -th sample. Let $\mathbf{I}_f^i = h_I(\mathbf{I}^i)$ and $\mathbf{T}_f^i = h_T(\mathbf{T}^i)$ denote the visual and the textual feature embeddings of the i -th sample, and $h_I(\cdot)$ and $h_T(\cdot)$ are the embedding functions. Assuming the textual feature embedding $h_T(\cdot)$ is known (e.g., using the embedding function of the report encoder in the ITM-branch), a report sample-graph could be constructed with a similarity matrix $\mathbf{W}_T$ whose element $\mathbf{W}_T^{ij} = \exp(\mathbf{T}_f^i \cdot \mathbf{T}_f^j / \sigma)$. The ground-truth report sample-graph serves as the target to train the image-based graph, and consequently the image feature embedding $h_I(\cdot)$. More specifically, the similarity matrix $\mathbf{W}_I$ for the image sample-graph is computed by: $\mathbf{W}_I^{ij} = \exp(\mathbf{I}_f^i \cdot \mathbf{I}_f^j / \sigma) = \exp(h_I(\mathbf{I}^i) \cdot h_I(\mathbf{I}^j) / \sigma)$. The two graphs $\mathbf{W}_I$ and $\mathbf{W}_T$ are normalized so that each row of $\mathbf{W}_I$ and $\mathbf{W}_T$ sums to 1. The embedding function $h_I(\cdot)$ could be trained by minimizing the cross-entropy between the two normalized graphs $\hat{\mathbf{W}}_I$ and $\hat{\mathbf{W}}_T$:

$$\mathcal{L}_{graph} = - \sum_{i=1}^B \sum_{j=1}^B \hat{\mathbf{W}}_I^{ij} \log(\hat{\mathbf{W}}_T^{ij}) \quad (4.6)$$

where $\hat{\mathbf{W}}_T^{ij} = \mathbf{W}_T^{ij} / \sum_{j=1}^B \mathbf{W}_T^{ij}$ and $\hat{\mathbf{W}}_I^{ij} = \mathbf{W}_I^{ij} / \sum_{j=1}^B \mathbf{W}_I^{ij}$, so that the structure difference between the two graphs is minimized. The sample-graph consistency loss is imposed as another regularization of the RG-branch for report generation.

4.3 Experiments

4.3.1 Datasets

In the experiment, two datasets are used for the performance evaluation. One (IU-Xray) represents the most widely used benchmark for medical report generation, while the other (MIMIC-CXR) is a recently released largest dataset to date.

IU-Xray Indiana University Chest X-ray Collection (IU-Xray) (Demner-Fushman *et al.*, 2016) is the most widely used publicly accessible dataset in medical report generation task. It contains 3955 fully de-identified radiology reports, each of which is associated with a frontal and/or lateral chest X-ray images, and 7470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings and Indication etc. In this work, we filtered out reports without two complete image views or without complete sections of “findings” and “impression”, resulting in a smaller dataset with 3195 reports associated with 6390 images. We tokenized all the words in “findings” and “impression” and obtained 2076 unique words. In addition, two special tokens, $\langle start \rangle$ and $\langle end \rangle$, are added to indicate the start and the end of a sentence. We randomly picked 311 (10%) reports to form the test set, which is comparable to the works in the literature (Jing *et al.*, 2018; Xue *et al.*, 2018). All evaluations are done on the test set.

MIMIC-CXR The recently released MIMIC-CXR (Johnson *et al.*, 2019) is the largest public dataset containing both the chest radiographs and free-text reports. In total, it consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016. In our experiment, we adopt MIMIC-CXR’s official split following the work (Chen *et al.*, 2020b) for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test, respectively. We convert all tokens to lowercase characters and remove non-word illegal characters and the words whose occurring frequency lower than 10, counting to 4030 unique words remaining in the dataset.

4.3.2 Experimental Settings

Evaluation Metrics To evaluate the quality of the generated text report, we utilize the widely used BLEU scores (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam *et al.*, 2015) as

evaluation metric. We compute those metrics by the standard image captioning evaluation tool².

Implementation Details We use the paired data as the input for IU-Xray dataset and concatenate the features of frontal and lateral images. The margin parameter α in Eqn. 4.2 and 4.3 is set to 0.2 and $\lambda_{CE}, \lambda_{feat}, \lambda_{SGCL}, \lambda_{ITM} = 2, 10, 2.2, 1$ in Eqn. 4.4, respectively. The word embedding size and topic vector size were respectively set to 300 and 512. We train our model using Adam optimizer (Kingma and Ba, 2015) with mini-batch size of 16. The learning rate are set to be 0.0001 and 0.0002, respectively, for RG-branch and ITM-branch with a total 30 epochs. We set $k = 10$ in Eqn. 4.3, training ITM branch with the regular triplet loss for the first 10 epochs and the self-boosted triple loss for the rest.

4.3.3 Results and Discussion

Comparison with SOTA We compare our model with 7 state-of-the-art image captioning methods. These include the classic model Show-tell (Vinyals *et al.*, 2015), different attention-based methods including AdaAtt (Lu *et al.*, 2017), Att2in (Anderson *et al.*, 2018), and Topdown (Rennie *et al.*, 2017), methods using advanced NLP models including Transformer (Devlin *et al.*, 2019) and M2transformer (Cornia *et al.*, 2020), and the very recent method Grounded (Zhou *et al.*, 2020) that also takes advantage of the image-text matching task to improve image captioning but in a completely different manner as ours. For these methods, we download the codes released publicly and re-run them on these two datasets with the same experiment setting as ours. Moreover, we also compare with 8 state-of-the-art medical report generation models: CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et al.*, 2018), KERP (Li *et al.*, 2019a), R2Gen (Chen *et al.*, 2020b), R2GenCMN (Chen *et al.*, 2021), PP-KED (Liu *et al.*, 2021a), AlignTransformer (You *et al.*, 2021) and CSK (Yang *et al.*, 2021).

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

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
IU-Xray	Show-Tell	0.346	0.214	0.141	0.095	0.320	0.161	0.239
	Att2in	0.399	0.249	0.172	0.126	0.321	0.165	0.341
	AdaAtt	0.436	0.288	0.203	0.150	0.354	0.168	0.265
	TopDown	0.375	0.237	0.167	0.123	0.319	0.164	0.336
	Transformer	0.422	0.264	0.177	0.120	0.338	0.163	0.268
	M2Transformer	0.463	0.318	0.214	0.155	0.335	0.171	0.349
	Grounded	0.446	0.301	0.237	0.176	0.343	0.176	0.395
	X-Transformer	0.467	0.322	0.235	0.169	0.351	0.178	0.393
	ViTCAP	0.474	0.329	0.248	0.181	0.376	0.184	0.401
	CoAtt [†]	0.455	0.288	0.205	0.154	0.369	-	0.277
	HGRG-Agent [†]	0.438	0.298	0.208	0.151	0.322	-	0.343
	KERP [†]	0.482	0.325	0.226	0.162	0.339	-	0.280
	R2Gen	0.445	0.283	0.202	0.155	0.362	0.173	0.353
	R2GenCMN	0.451	0.292	0.208	0.159	0.363	0.177	0.374
	PPKED [†]	0.483	0.315	0.224	0.168	0.376	0.187	0.351
	AlignTransformer [†]	0.484	0.313	0.225	0.173	0.379	0.204	-
	GSK [†]	0.496	0.327	0.238	0.178	0.381	-	0.382
	Ours(HLSTM)	0.483	0.348	0.272	0.210	0.362	0.193	0.457
	Ours(MATM)	0.489	0.356	0.281	0.217	0.364	0.197	0.459
MIMIC-CXR	Show-Tell	0.308	0.190	0.125	0.088	0.256	0.122	0.096
	Att2in	0.314	0.198	0.133	0.095	0.264	0.122	0.106
	AdaAtt	0.314	0.198	0.132	0.094	0.267	0.128	0.131
	Transformer	0.316	0.199	0.140	0.092	0.267	0.129	0.134
	M2Transformer	0.332	0.210	0.142	0.101	0.264	0.134	0.142
	Grounded	0.352	0.219	0.147	0.105	0.274	0.144	0.167
	X-Transformer	0.356	0.221	0.148	0.109	0.277	0.148	0.185
	ViTCAP	0.367	0.225	0.152	0.113	0.282	0.156	0.233
	R2Gen	0.353	0.218	0.145	0.103	0.277	0.142	0.141
	R2GenCMN	0.353	0.218	0.148	0.106	0.278	0.142	0.143
	PPKED [†]	0.36	0.224	0.149	0.106	0.284	0.149	0.237
	AlignTransformer [†]	0.378	0.235	0.156	0.112	0.284	0.158	-
	GSK [†]	0.363	0.228	0.156	0.115	0.284	-	0.203
	Ours(HLSTM)	0.376	0.235	0.157	0.119	0.287	0.159	0.294
	Ours(MATM)	0.387	0.251	0.177	0.129	0.291	0.164	0.342

Table 4.1. Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). [†] indicates the results are quoted from the published literature. Specifically, we quote the results of CoAtt (Jing *et al.*, 2018) and HGRG-Agent (Li *et al.*, 2018) for IU-Xray dataset from (Li *et al.*, 2019a) while other methods we cite results from their respective papers. For methods without dagger tags in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method.

al., 2021b). Among these methods, R2Gen (Chen *et al.*, 2020b) and R2GenCMN (Chen *et al.*, 2021) have officially released codes, and we re-run them with the same experiment setting as ours since the CIDEr score was not provided in the original paper. For other models which did not release the codes, we quote the results of CoAtt (Jing *et al.*, 2018) and HGRG-Agent (Li *et al.*, 2018) for IU-Xray dataset from (Li *et al.*, 2019a) while other methods we cite results from their respective papers. Please note that since IU-Xray dataset has no official training-test

partition, the performances on IU-Xray cited from the original papers may not be strictly comparable and are provided here only for reference. Differently, on MIMIC-CXR, since PPKED (Liu *et al.*, 2021a), AlignTransformer (You *et al.*, 2021) and CSK (Yang *et al.*, 2021b) follow the MIMIC-CXR official training-test partition, their cited performances are comparable.

As shown in Table 4.1, on both datasets, our self-boosting method achieves the best performance over almost all evaluation metrics among the comparing methods. The two very recent image captioning methods M2transformer (Cornia *et al.*, 2020) and Grounded (Zhou *et al.*, 2020) perform better than other image captioning methods, but still lose to ours with either HLSTM decoder or MATM decoder. Especially, our advantages over Grounded (Zhou *et al.*, 2020) fully demonstrate the benefit to jointly train the two tasks of report generation and image-text matching in our proposed self-boosting manner. It is also noted that, our performance improvement on CIDEr is most significant. Cross-referring to Table 4.4, even without the reinforcement learning on CIDEr (denoted as Ours-RL), our method is still the best performer on CIDEr on both datasets and the advantage is pronounced³. This is a positive sign for radiographic reports, since CIDEr focuses on key information by down-weighting the common words in all reports. Moreover, both Ours(HLSTM) and Ours (MATM) win three recent medical report generation methods PPKED (Liu *et al.*, 2021a), AlignTransformer (You *et al.*, 2021) and CSK (Yang *et al.*, 2021b) on MIMIC-CXR in terms of all the evaluation metrics in Table 4.1. Ours (MATM) also achieves better results (except ROUGE) than these methods on IU-Xray, while it is again worth mentioning that the cited performances on IU-Xray may not be strictly comparable as obscure training-test partitions were used by different methods. All these three methods explore and build additional knowledge graph to enhance radiographic report generation, while ours have not taken advantage

³It is noted that Grounded (Zhou *et al.*, 2020) also uses RL on CIDEr for refinement.

	Ground-truth	M2transformer	Grounded
	Impression: No acute cardiopulmonary process. Findings: Coronary artery stents visualized. The cardio mediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. No acute bone abnormality.	Impression: No acute cardiopulmonary process. Findings: The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion. cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	Impression: No acute cardiopulmonary abnormality. Findings: The cardio mediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear. there is no pleural effusion or pneumothorax. No acute bone abnormality.
	R2Gen	RG-branch(CE only)	Ours
	Impression: No acute cardiopulmonary process. Findings: The cardio mediastinal silhouette, pulmonary vascular pattern are within normal limits. The lungs are clear of focal airspace disease pneumothorax or pleural effusion, or pneumothorax. Mild degenerative changes mid thoracic spine.	Impression: No acute cardiopulmonary process. Findings: The cardio mediastinal silhouette is within normal limits for size and contour. The lungs are clear. there is no pleural effusion or pneumothorax. the osseous structures are intact. No subdiaphragmatic free air.	Impression: No acute cardiopulmonary process. Findings: The cardio mediastinal silhouette and pulmonary vasculature are within normal limits for size and contour. The lungs are clear of focal airspace disease pneumothorax or pleural effusion, or pneumothorax. there are no acute bony abnormality.

Figure 4.4. Examples of the generated reports on IU-xray dataset. We compare our results with the ground truth report, our baseline using RG-branch only with the cross-entropy loss (denoted as RG-branch(CE only)), and another three competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports represent incorrectly predicted sentences, sentences not exactly consistent with ground-truth, and sentences that completely match the ground-truth, respectively.

of such additional information yet. This further demonstrates the effectiveness of the proposed self-boosting framework.

In addition, we conduct qualitative comparison between ours (MATM) and three comparing methods: the image captioning methods M2transformer (Cornia *et al.*, 2020) and Grounded (Zhou *et al.*, 2020) and the medical report generation method R2Gen (Chen *et al.*, 2020b). These three methods are recently published with released codes. The examples of the generated reports are visually compared in Figure 4.4 and Figure 4.5 for IU-Xray dataset and MIMIC-CXR dataset, respectively. To facilitate comparison, we highlighted the sentences with red, blue and green color in the generated reports to represent incorrectly predicted sentences, sentences not exactly consistent with ground-truth, and sentences that completely match the ground-truth, respectively. As can be observed, compared with our baseline (RG branch only) and the other three competitive

	Ground-truth	M2transformer	Grounded
	Findings: As compared to previous radiograph the right pleural effusion has slightly decreased. On the left however a relatively extensive pleural effusion occupying approximately one-third of the left hemithorax has newly appeared. Subsequent areas of atelectasis. No pneumothorax or pneumonia. No evidence of pulmonary edema.	Findings: Frontal and lateral views of the chest were obtained. The heart size is normal. the mediastinal and hilar contours are normal. the lungs are hyperinflated. no focal consolidation pleural effusion or pneumothorax is seen. there is no pulmonary edema.	Findings: pa and lateral views of the chest. the lungs are hyperinflated. there is no focal consolidation pleural effusion or pneumothorax. the heart size is normal . the mediastinal and hilar contours are normal. there is no pulmonary edema.
	R2Gen	RG-branch(CE only)	Ours
	Findings: as compared to the previous radiograph the tip of the nasogastric tube is in the stomach. the tip of the nasogastric tube is in the stomach. the heart size is normal . the mediastinal and hilar contours are normal. the lungs are clear . subsequent areas of atelectasis are bilaterally constant. there is no pleural effusion or pneumothorax. there is no pulmonary edema.	Findings: As compared to the previous radiograph there is no relevant change . moderate bilateral pleural effusions with subsequent areas of atelectasis at both lung bases . no new parenchymal opacities . unchanged size of the cardiac silhouette . No pneumothorax or pneumonia. There is no evidence of pulmonary edema.	Findings: as compared to the previous radiograph the right chest tube has been removed . there is a small left pleural effusion with subsequent areas of atelectasis at the left lung base. the right lung is unchanged . No pneumothorax or pneumonia. There is no evidence of pulmonary edema.

Figure 4.5. Examples of the generated reports on MIMIC-CXR dataset. We compare our results with the ground truth report, our baseline using RG-branch only with the cross-entropy loss (denoted as RG-branch(CE only)), and another three competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports represent incorrectly predicted sentences, sentences not exactly consistent with ground-truth, and sentences that completely match the ground-truth, respectively.

methods, our proposed model can significantly improve the quality of the generated reports, i.e., producing more sentences in green and less sentences in red, consistent with the above quantitative analysis.

We also report the model’s training time and capacity in Table 4.2. Specifically, we trained the models on two Nvidia 3090 GPUs in a total of 30 epochs with a batch size of 16. Show-tell, Att2in, TopDown, and Transformer are reproduced by this code base ⁴, R2Gen and R2GenCMN are reproduced from their respective publicly available code bases. As can be seen, compare with R2Gen (Chen *et al.*, 2020b), our model has a smaller number of parameters (790.9 vs 1200) and higher training efficiency (1.65h vs 1.79h), while achieving better performance.

Evaluation on Image Captioning To investigate the generalization ability of our model, we further evaluate the performance of our model on the general

⁴<https://github.com/ruotianluo/self-critical.pytorch>

Table 4.2. Model running time and capacity trained on IU-Xray dataset.

Method	Running Time	Model Capacity	BLEU_4	CIDEr
Show-Tell	1.15h	51.1M	0.095	0.239
Att2in	1.26h	55.1M	0.126	0.341
TopDown	1.35h	198.8M	0.123	0.336
Transformer	1.58h	219.3M	0.120	0.268
R2Gen	1.79h	1200M	0.155	0.353
R2GenCMN	1.82h	978M	0.159	0.374
Ours(HLSTM)	1.32h	485.5M	0.210	0.457
Ours(MATM)	1.65h	790.9M	0.217	0.459

image captioning task. We conduct all the experiments on the most popular image captioning benchmark COCO (Lin *et al.*, 2014), and we adopt the widely adopted Karpathy split (Karpathy and Fei-Fei, 2015) for offline evaluation. There are 113287 training images, 5000 validation images, and 5000 testing images in the Karpathy split. Following (Pan *et al.*, 2020), we pre-process all training sentences by converting them into lower case and dropping the words that occur less than 6 times, leading to the final vocabulary with 9488 unique words.

For a fair comparison, we take a SOTA image captioning model X-Transformer (Pan *et al.*, 2020) as the backbone model and incorporate our proposed self-boosting and sample-graph consistency strategies into X-Transformer framework to investigate its effectiveness. As shown in Table 4.3, our enhanced approach significantly improves the performance of the baseline X-Transformer model. Specifically, the BLEU-4 score increases from 0.370 to 0.376, the METEOR score from 0.287 to 0.290, the ROUGE score from 0.575 to 0.581, and the CIDEr score from 1.200 to 1.226. We compare Ours(X-Transformer) with 7 state-of-the-art general image captioning methods, including TopDown (Anderson *et al.*, 2018), LBPF (Qin *et al.*, 2019), SGAE (Yang *et al.*, 2019), SGAE-KD (Yang *et al.*, 2020b), AoANet (Huang *et al.*, 2019a) and two latest models TopDown+DCR (Wang *et al.*, 2022a) and ViTCAP (Fang *et al.*, 2022) (without pretraining for a fair comparison). As can be seen, Ours(X-Transformer)

achieves the best performance over all evaluation metrics among the methods in comparison. It should be noted that pre-training-based methods like VinVL (Zhang *et al.*, 2021) outperform our X-Transformer model with scores of 0.382 for Bleu4, 0.303 for Rouge, and 1.293 for CIDEr. This superior performance is attributed to their use of extensive external datasets for pre-training. However, these methods require extensive data and are computationally intensive, resulting in time-consuming training.

Table 4.3. Performance comparisons on COCO Karpathy test split.

Methods	Conf&Year	BLEU-4	METEOR	ROUGE	CIDEr
TopDown	CVPR2018	0.359	0.275	0.562	1.113
TopDown+DCR	TPAMI2022	0.363	0.278	0.565	1.114
LBPF	CVPR2019	0.374	0.281	0.575	1.164
SGAE	CVPR2019	0.369	0.277	0.572	1.167
SGAE-KD	TPAMI2020	0.373	0.281	0.574	1.171
AoANet	ICCV2019	0.372	0.284	0.575	1.198
ViTCAP	CVPR2022	0.357	0.288	0.576	1.218
X-Transformers	CVPR2020	0.370	0.287	0.575	1.200
Ours(X-Transformers)	-	0.376	0.290	0.581	1.226

Ablation Study We conduct an ablation study to single out the contributions of each component in our proposed method, as presented in Table 4.4. We take the RG branch (with only the traditional cross-entropy loss) of our proposed pipeline as the baseline to verify the performance improvements brought by ITM-branch and our self-boosted training strategy. In Table 4.4, there are five components: ITM, FLL, STL, RL and SGCL, representing Image-Text matching, Feature-Level Loss, Self-boosted Loss, Reinforcement Learning, and Sample Graph Consistency Loss, respectively. The symbols “+” or “-” indicate the inclusion or exclusion of the following component. The benefit of using the auxiliary image-text matching task can be well reflected by the improvement from “RG-branch(CE only)” to “RG-branch(CE only)+ITM”. As shown, the performance can be further boosted by additionally utilizing feature-level loss (“RG-branch(CE only)+ITM+FLL”) in RG-branch and the self-boosted triplet

Dataset	Model	BLEU-4	ROUGE	METEOR	CIDEr
IU-Xray	RG-branch(CE only)	0.146	0.319	0.171	0.365
	RG-branch(CE only)+ITM	0.162	0.330	0.178	0.387
	RG-branch(CE only)+ITM+FLL	0.170	0.334	0.181	0.396
	RG-branch(CE only)+ITM+STL	0.184	0.348	0.184	0.424
	RG-branch(CE only)+ITM+STL+SGCL	0.197	0.356	0.187	0.433
	Ours-RL	0.203	0.361	0.189	0.438
	Ours	0.217	0.364	0.197	0.459
MIMIC-CXR	RG-branch(CE only)	0.103	0.267	0.143	0.169
	RG-branch(CE only)+ITM	0.106	0.271	0.147	0.193
	RG-branch(CE only)+ITM+FLL	0.108	0.272	0.150	0.205
	RG-branch(CE only)+ITM+STL	0.112	0.276	0.153	0.231
	RG-branch(CE only)+ITM+STL+SGCL	0.116	0.282	0.157	0.245
	Ours-RL	0.118	0.284	0.158	0.253
	Ours	0.129	0.291	0.164	0.342

Table 4.4. Ablation studies. We use our RG-branch only with the traditional cross entropy-loss as our baseline model for ablation experiments, which is denoted as “RG-branch(CE only)”. Here “CE”, “ITM”, “FLL”, “STL”, “RL”, and “SGCL” stand for “Cross-entropy Loss”, “Image-Text Matching”, “Feature Level Loss” ($\mathcal{L}_{feat}$ in Eqn. 4.1), “Self-boosted Triplet Loss” (Eqn. 4.3), “Reinforcement Learning”, and “Sample Graph Consistency Loss” ($\mathcal{L}_{graph}$ in Eqn. 4.1), respectively.

Table 4.5. Top 10 Anatomy entities and Observation entities with the highest frequency in Radgraph (Jain *et al.*, 2021)

Top-10 Anatomy entities	Top-10 Observation entities
pleural	normal
lung	unchanged
pulmonary	clear
silhouette	atelectasis
contours	effusion
mediastinal	opacity
lobe	mild
cardiac	edema
hilar	cardiomegaly
heart	unremarkable

loss (“RG-branch(CE only)+ITM+STL”) in ITM-branch, proving their indispensability in our proposed method. In addition, as can be seen, using our sample-graph consistency loss (“RG-branch(CE only)+ITM+STL+SGCL”), our model could further elevate the performance in all three metrics. Moreover, as mentioned, even removing the refinement by reinforcement learning, our model still significantly outperforms the comparing methods in Table 4.1.

Hyper-Parameters study We conducted a sensitivity analysis on hyperparameters on the MIIMC-CXR dataset. First, we fixed the $\lambda_{ITM}=1$, $\lambda_{SGCL}=0$ and increasing/decreasing λ_{CE} , λ_{feat} by 1.5 times to evaluate the sensitivity of the model to the first three hyperparameters λ_{CE} , λ_{feat} and λ_{ITM} . As can be observed from Table 4.6, when λ_{CE} and λ_{feat} vary over a reasonably wide range, the performance of the model does not show significant fluctuations. Then, we fixed three hyperparameters: $\lambda_{CE}=2$, $\lambda_{feat}=10$, $\lambda_{ITM}=1$, and increased and decreased our newly introduced hyperparameter λ_{SGCL} by 25% and 50%, respectively. The experimental results are shown in Table 4.7. We observed that the model is not sensitive to changes in the parameter when it is increased and decreased by 25%, with fluctuations of 0.7% and -3.1% for BLEU_4 and -5.8% and -1.2% for CIDEr, respectively. However, when the parameter is decreased by 50%, there is a significant drop in model performance. This may be due to the fact that the weight of $\mathcal{L}_{graph}$ is set too small, leading to insufficient training of the Sample-graph consistency loss.

Table 4.6. Hyper-parameter study using MIMIC-CXR (fix $\lambda_{ITM} = 1$ and $\lambda_{SGCL} = 0$)

λ_{CE}	λ_{feat}	BLEU-4	Rouge	METEOR	CIDEr
2.0	10.0	0.121	0.285	0.160	0.290
1.3 (2 / 1.5)	10.0	0.117	0.280	0.156	0.285
3.0 (2 $\times$ 1.5)	10.0	0.122	0.283	0.160	0.288
2.0	6.7 (10 / 1.5)	0.120	0.283	0.158	0.289
2.0	15.0 (10 $\times$ 1.5)	0.119	0.281	0.159	0.287

Table 4.7. Hyper-parameter study using MIMIC-CXR (fix $\lambda_{ITM} = 1$, $\lambda_{CE} = 2$ and $\lambda_{feat} = 10$)

λ_{SGCL}	BLEU-4	ROUGE	METEOR	CIDEr
2.2	0.129	0.291	0.164	0.342
1.10 (2.2 $\times$ 0.5)	0.122	0.286	0.160	0.292
1.65 (2.2 $\times$ 0.75)	0.125	0.288	0.161	0.322
2.75 (2.2 $\times$ 1.25)	0.130	0.290	0.163	0.338
2.00 (2.2 $\times$ 1.5)	0.127	0.289	0.162	0.334

We conducted further analysis on the impact of parameter k on model performance. We set k to four different values: 0, 10, 20, and 30. Specifically, when

$k=0$, it indicates that we do not use $L_{match-gen}$, while $k=30$ means that we only use $L_{match-gen}$. By examining Table 4.8, we can see that setting k to 30 (training solely with $L_{match-gen}$ from the start) results in a significant decrease in model performance, emphasizing the importance of preheating with L_{match} . On the other hand, when k is set to 0 (training solely with L_{match}), compared with training with both L_{match} and $L_{match-gen}$, the model performance is affected, which validates the effectiveness of our proposed matching strategy.

Table 4.8. Performance using different parameter k on MIMIC-CXR dataset.

K (epoch=30)	BLEU-4	METEOR	ROUGE	CIDEr
0	0.124	0.283	0.159	0.324
20	0.127	0.290	0.162	0.336
30	0.120	0.279	0.155	0.312
10	0.129	0.291	0.164	0.342

We further explore the impact of the alternate stride m on model performance on the MIMIC-CXR dataset. Specifically, we set m to 1, 3, 5, and 7, respectively, and train the model for a total of 30 epochs, with the original value of 5 epochs serving as a reference point. As observed in Table 4.9, when we set m to 1, the model performance shows a significant decline. This could be attributed to the unstable model training caused by the high switching frequency of RG branch and ITM branch, which prevents either branch of the model from being fully trained. Increasing the stride M to 3 significantly improved the model’s performance, which supports our hypothesis. Notably, these results were consistent with the performance achieved with a stride of 5, which was the default value used in our original study. However, setting m to 7 resulted in a slight decrease in performance, possibly due to insufficient interaction between the two branches.

Clinical accuracy evaluation In addition to the most commonly used metrics, like BLEU and CIDEr, we further provide Keyword Accuracy (KA) in Table 4.10 by following the works (Xue *et al.*, 2018; Huang *et al.*, 2019b) to evaluate the Clinical Correctness of the generated reports. KA is the ratio of

Table 4.9. Performance using different alternate stride on MIMIC-CXR dataset.

M (epoch=30)	BLEU-4	METEOR	ROUGE	CIDEr
1	0.114	0.283	0.157	0.279
3	0.126	0.292	0.163	0.343
7	0.122	0.288	0.161	0.334
5	0.129	0.291	0.164	0.342

Table 4.10. Clinical accuracy evaluation. Keyword Accuracy (KA) is used to evaluate Clinical Correctness of the generated report, which is the ratio of the number of correctly generated keywords to the number of all keywords in the ground truth findings.

Dataset	Model	Keyword Accuracy
MIMIC-CXR	M2Transformer	0.863
	Grounded	0.870
	R2Gen	0.864
	R2GenCMN	0.869
	Ours(HLSTM)	0.884
	Ours(MATM)	0.890

the number of correctly generated keywords to the number of all keywords in the ground truth findings. We compare our model with two image captioning methods M2Transformer (Cornia *et al.*, 2020) and Grounded (Zhou *et al.*, 2020), and two medical report generation models R2Gen (Chen *et al.*, 2020b) and R2GenCMN (Chen *et al.*, 2021), all of which are recently published with released codes. Since the works (Xue *et al.*, 2018; Huang *et al.*, 2019b) did not release their keywords or the selection criteria, we constructed a keyword dictionary with 768 keywords based on Radgraph’s (Jain *et al.*, 2021) high frequency entities. Radgraph contains carefully selected entities annotated by three board-certified radiologists, including the Anatomy entities refers to an anatomical body part that occurs in a radiology report, such as a “lung”, and the Observation entities which are associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications, such as “effusion”. We list the Top-10 Anatomy entities and Observation entities with the highest frequency in Radgraph in Table 4.5 for reference. As seen in Table 4.10,

both Ours(HLSTM) and Ours(MATM) achieve significantly better KA values, suggesting high fidelity of our models in clinic correctness.

4.4 Conclusions

In this work, we improve the diagnostic report generation by additionally utilizing an auxiliary image-text matching task to learn strongly correlated visual and text features to describe fine-grained differences among radiographic images. We show how to achieve this by deeply coupling the two tasks and encouraging the cooperative interactions between them. By communicating the learned features and the newly generated samples, the two tasks could mutually boost each other in a progressive way. Moreover, we further improve the report generation task by minimizing the losses of feature-level similarity and sample-graph consistency. The former requires the generated reports to stay similar to the ground-truth reports at high semantic levels. The latter requires the sample relationships remain similar between the report samples and the image samples in the feature embedding space, and uses the sample relationship of the ground-truth reports to guide the learning of the visual features. As experimentally verified, our generated reports could better capture the subtle but critical information in radiographic images.

A Medical Semantic-Assisted Transformer for Radiographic Report Generation

Automated radiographic report generation is a challenging cross-domain task that aims to automatically generate accurate and semantic-coherence reports to describe medical images. Despite the recent progress in this field, there are still many challenges at least in the following aspects. First, radiographic images are very similar to each other, and thus it is difficult to capture the fine-grained visual differences using CNN as the visual feature extractor like many existing methods. Further, semantic information has been widely applied to boost the performance of generation tasks (e.g. image captioning), but existing methods often fail to provide effective medical semantic features. Toward solving those problems, in this chapter, we propose a memory-augmented sparse attention block utilizing bilinear pooling to capture the higher-order interactions between the input fine-grained image features while producing sparse attention. Moreover, we introduce a novel Medical Concepts Generation Network (MCGN) to predict fine-grained semantic concepts and incorporate them into the report generation process as guidance. Our proposed method shows promising performance on the recently released largest benchmark MIMIC-CXR. It outperforms multiple state-of-the-art methods in image captioning and medical report generation.

5.1 Introduction

While significant progress has been made in the field of automated radiographic report generation, as discussed in previous chapters, nuanced challenges specific to this domain persist. First, unlike natural images, radiographic images are very similar to each other. The medical report generation methods using the CNN-based image encoder extract image-level features due to the lack of image region annotations in such applications, which could not well cater for the local details reflecting the fine-grained image patterns of clinic importance. This problem could be somewhat mitigated by employing the recently developed Vision Transformer model (Dosovitskiy *et al.*, 2021) that explores the dependency of image regions without generating regional proposals. However, there still lacks mechanism to better treat fine-grained patterns. Second, the differences between medical reports are also fine-grained, that is, the reports are dominated by similar sentences describing the common content of the images while the disease-related words may be submerged. A possible remedy is to incorporate semantic textual concepts into the model training to guide the report generation. However, such information has not been well explored for medical report generation, especially in the recently popular transformer-structured models.

To cope with the limitations mentioned above, we propose the following solutions to advance radiographic report generation. Firstly, we carefully deal with the fine-grained differences existed in radiographic images from two aspects. On the one hand, we utilize CLIP (Radford *et al.*, 2021) rather than ResNet as our visual feature extractor. The advantage of CLIP is two-fold: i) it is built upon Vision Transformer that extracts regional visual features and relationships, and ii) it is trained by matching image-text pairs and thus produces text-enhanced image features. On the other hand, more importantly, we introduce an memory-augmented sparse attention for high-order feature interactions and embed it into the transformer encoders and decoders. This attention makes use of bilinear pooling that proves to be effective for fine-grained visual recognition (Lin *et al.*,

2015), memorizes historical information for long report generation, and produces sparse attention for efficient report generation.

Second, to incorporate semantic textual concepts, we introduce a novel lightweight medical concepts generation network to predict fine-grained semantic concepts and incorporate them into medical report generation process, which is different from the usage of sparse medical tags in the existing methods (Jing *et al.*, 2018). In sum, our main contributions include: 1) We introduce a bilinear-pooling-assisted sparse attention block and embed it into a transformer network to capture the fine-grained visual difference existed between radiographic images; 2) We propose a medical concepts generation network to provide enriched semantic information to benefit radiographic report generation; 3) We extensively validate our model on the recently released largest dataset MIMIC-CXR.

5.2 Method

The proposed framework adopts the encoder-decoder paradigm, where the transformer encoder and decoder are embedded with a Memory-augmented Sparse Attention block (MSA) to capture the higher-order interactions between the input fine-grained image features. Meanwhile, a Medical semantic Concepts Generation Network (MCGN) is incorporated into the report generation model to further improve the performance. In the following, we first introduce the MSA block in Section 5.2.1 and how the MSA block are embedded into the encoder and decoder in Section 5.2.2 and Section 5.2.4, and our proposed MCGN in Section 5.2.3.

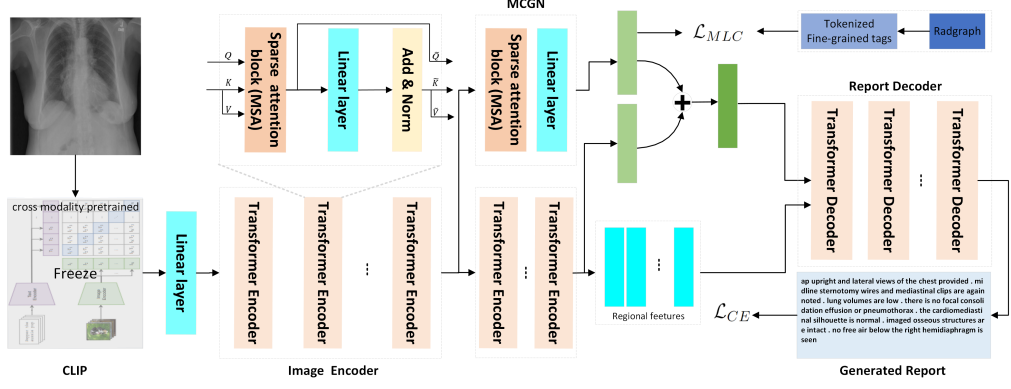


Figure 5.1. An overview of the proposed framework, which comprises an Image Encoder, a Medical Concepts Generation Network (MCGN) and a Report Decoder. The transformer encoder and decoder are embedded with a Memory-augmented Sparse non-linear Attention (MSA) to capture the higher-order interactions between the input fine-grained image features. Meanwhile, we inject a medical concepts generator to provide semantic information to facilitate report generation.

5.2.1 Memory-augmented Sparse Attention of High-order Interaction

Given the three input matrices, queries $\mathbf{Q} \in \mathbf{R}^{N \times D_q}$, keys $\mathbf{K} \in \mathbf{R}^{N \times D_k}$ and values $\mathbf{V} \in \mathbf{R}^{N \times D_v}$, the conventional self-attention block (Vaswani *et al.*, 2017) employed by transformer first computes the dot products (element-wise sum) of each query with all keys and then applies a softmax function to obtain the weights on the values. Formally, the attention is calculated by:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (5.1)$$

It exploits only 1st order interaction of the input vectors since the weight is derived from the linear fusion of the given query and keys via element-wise sum. However, the linear attention may not be sufficient to capture the fine-grained visual differences among the input radiographic images. Inspired by the recent success of bilinear pooling applied in fine-grained visual recognition, we inject bilinear-pooling into the self-attention to capture the 2nd or even higher-order (by stacking these blocks) interactions of the input fine-grained visual features.

Specifically, we assume the global feature as the query $\mathbf{Q} \in \mathbf{R}^{D_q}$, the regional features as the key $\mathbf{K} \in \mathbf{R}^{N \times D_k}$ and value $\mathbf{V} \in \mathbf{R}^{N \times D_v}$. To record the historical information, we extent the set of keys and values with additional “memory-slots” to encode and collect the features from all the previous processes. The key and value of our memory-augment attention can be defined as: $\hat{\mathbf{K}} = [\mathbf{K}, \mathbf{M}_k]$ and $\hat{\mathbf{V}} = [\mathbf{V}, \mathbf{M}_v]$, respectively, where $\mathbf{M}_k$ and $\mathbf{M}_v$ are learnable matrices with n_m rows, and $[\cdot, \cdot]$ indicates concatenation. Then, a low-rank bilinear pooling (Kim *et al.*, 2017) is performed to obtain the joint bilinear query-key $\mathbf{B}_k$ and query-value $\mathbf{B}_v$ by

$$\mathbf{B}_k = \sigma(\mathbf{W}_k \hat{\mathbf{K}}) \odot \sigma(\mathbf{W}_q^k \mathbf{Q}), \quad \mathbf{B}_v = \sigma(\mathbf{W}_v \hat{\mathbf{V}}) \odot \sigma(\mathbf{W}_q^v \mathbf{Q}) \quad (5.2)$$

where $\mathbf{W}_k \in \mathbf{R}^{D_B \times D_k}$, $\mathbf{W}_v \in \mathbf{R}^{D_B \times D_v}$, $\mathbf{W}_q^k \in \mathbf{R}^{D_B \times D_q}$ and $\mathbf{W}_q^v \in \mathbf{R}^{D_B \times D_v}$ are learning parameters, σ denotes ReLU unit, and $\odot$ represents element-wise multiplication. Next, we use a linear layer to project $\mathbf{B}_k \in \mathbf{R}^{D_c \times D_B}$ into a intermediate representation $\hat{\mathbf{B}}_k = \sigma(\mathbf{W}_B^k \mathbf{B}_k)$, then use another linear layer to map $\hat{\mathbf{B}}_k$ from D_c dimension to 1 dimension to obtain the spatial-wise attention weight $\alpha_s \in \mathbf{R}^{D_c \times 1}$. Unlike (Pan *et al.*, 2020) using softmax to normalize α_s , we utilized another ReLU unit to prune out all negative scores of low query-key relevance, automatically ensuring the sparse property of the attention weight $\beta_s = \sigma(\alpha_s)$. Meanwhile, we perform a squeeze-excitation operation (Hu *et al.*, 2018) to $\hat{\mathbf{B}}_k$ to obtain channel-wise attention $\beta_c = \text{sigmoid}(\mathbf{W}_c) \bar{\mathbf{B}}$, where $\mathbf{W}_c \in \mathbf{R}^{D_B \times D_c}$ is learnable parameters and $\bar{\mathbf{B}} \in \mathbf{R}^{D_B \times 1}$ is an average pooling of $\hat{\mathbf{B}}_k$. The output attended features of our memory-augmented sparse attention integrate the enhanced bilinear values with spatial and channel-wise bilinear attention $\hat{\mathbf{Q}} = \text{Attention}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \beta_c \odot \text{LN}(\beta_s \mathbf{B}_v)$. where $\text{LN}(\cdot)$ denotes variants of layer normalization (Ba *et al.*, 2016).

Such structure design benefits the model in three ways. First, it can explore higher-order interactions between the input single-model (in the encoder) or multi-model (in the decoder) features, resulting in a more robust representative

capacity of the output attended features. Second, the memory tokens we embed into the MSA block can record the previous generation process and collect historical information, which can be valuable for our task’s long report generation. Third, we propose a softmax-free sparse attention module, pruning out all negative attention scores and producing sparse attention of transformer decoder with higher efficiency in the process of report generation. We measured the speedup per training step on 500 steps with about 16 samples per batch to compare the running efficiency of softmax-base attention and our sparse Relu-based attention. We perform three runs on a single NVIDIA TESLA V100 and report average results. Without performance degradation, our sparse attention can reduce the training time from 0.497s per batch to 0.488s per batch.

5.2.2 Image Encoder

Let’s denote an input image by $\mathbf{I}$. The pre-trained CLIP model (ViT-B/16) is utilized to extract the regional features of $\mathbf{I}$: $f = \text{CLIP}(\mathbf{I})$, where $f \in \mathbb{R}^{D_c \times D_f}$. We take the mean-pooled feature embedding f as the initial $\mathbf{Q}^{(0)} = \frac{1}{N_c} \sum_{i=1}^{N_c} f_i$, and f as the initial $\mathbf{K}^{(0)}$ and $\mathbf{V}^{(0)}$, and feed them into the encoding layers.

Encoding Layer We embed our memory-augmented sparse attention into a Transformer-like layer. Formally, for the m -th transformer encoder TE_m , we take the previous output attended feature $\hat{\mathbf{Q}}^{(m-1)}$ as the input query, couple it with the current input keys $\mathbf{K}^{m-1} = \{\mathbf{k}_i^{(m-1)}\}_{i=1}^N$, and values $\mathbf{V}^{(m-1)} = \{\mathbf{v}_i^{m-1}\}_{i=1}^N$, which could be expressed as:

$$\mathbf{K}^{(m)}, \mathbf{V}^{(m)}, \mathbf{Q}^{(m)} = TE_m(\mathbf{K}^{(m-1)}, \mathbf{V}^{(m-1)}, \hat{\mathbf{Q}}^{(m-1)}) \quad (5.3)$$

where $\hat{\mathbf{Q}}^{(m)} = \text{Attention}(\mathbf{K}^{(m-1)}, \mathbf{V}^{(m-1)}, \hat{\mathbf{Q}}^{(m-1)})$ computed by MSA block(see sec. 5.2.1). Then, all the keys and values are further updated conditioned on the new attended feature $\hat{\mathbf{Q}}^{(m)}$ by equation: $\mathbf{k}_i^{(m)} = \text{LN}(\sigma(\mathbf{W}_m^k [\hat{\mathbf{Q}}^{(m)}, \mathbf{k}_i^{(m-1)}]) + \mathbf{k}_i^{(m-1)})$ and $\mathbf{v}_i^{(m)} = \text{LN}(\sigma(\mathbf{W}_m^v [\hat{\mathbf{Q}}^{(m)}, \mathbf{v}_i^{(m-1)}]) + \mathbf{v}_i^{(m-1)})$, respectively, where $\mathbf{W}_m^k$ and $\mathbf{W}_m^v$ are learnable parameters. $[\cdot, \cdot]$ indicates concatenation. This means

that each key or value is concatenated with the new attended feature, followed with a residual connection and layer normalization (Ba *et al.*, 2016). Noted that the entire image encoder consists of $M = 6$ encoding layers in this chapter. In particular, when $m = 1$, $\hat{\mathbf{Q}}^{(0)} = \mathbf{Q}^{(0)} = \frac{1}{N_c} \sum_{i=1}^{N_c} f_i$.

5.2.3 Medical Concepts Generation Network

In order to train the Medical Concepts Generation Network (MCGN), we utilize RadGraph (Jain *et al.*, 2021) to extract the pseudo-medical concepts as the ground-truth for multi-label classification (MLC). Specifically, the RadGraph is a knowledge graph of clinic radiology entitles and relations based on full-text chest x-ray radiology reports. We use these entities as the medical semantic concepts and finally select 768 concepts according to their occurrence frequency.

Our MCGN consists of an MSA block for processing the intermediate features generated by the m -th Encoding layer of Image Encoder. The output of MSA is $\mathbf{V}_c = \text{Attention}(\mathbf{K}^{(m-1)}, \mathbf{V}^{(m-1)}, \hat{\mathbf{Q}}^{(m-1)})$, where $m \in [1, M]$ denotes the m -th encoding layer. Then, a linear projection maps $\mathbf{V}_c$ from D to K dimensions, where K is the number of medical concepts. The overall medical concepts classification loss can be expressed as:

$$\begin{aligned} \mathcal{L}_{MLC} = & -\frac{1}{K} \cdot \sum_i y_i \cdot \log((1 + \exp(-x_i))^{-1}) \\ & + (1 - y_i) \cdot \log\left(\frac{\exp(-x_i)}{1 + \exp(-x_i)}\right), \end{aligned} \quad (5.4)$$

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

5.2.4 Report Decoder

The report decoder aims to generate the output report conditioned on the attended visual embeddings $\mathbf{V}_i = \hat{\mathbf{Q}}^{(m)}_{m=0}$ from image encoder and the medical concepts embeddings $\mathbf{V}_c$ from MCGN. Therefore, we first fused those embeddings by the equation: $\mathbf{v}_f = \mathbf{W}_f[\hat{\mathbf{Q}}^{(0)}, \hat{\mathbf{Q}}^{(1)}, \dots, \hat{\mathbf{Q}}^{(M)}] + \mathbf{V}_c$, where $\mathbf{W}_f$ is a learnable parameter. The input of transformer decoder is thus set as the concatenation of the fused feature $\mathbf{v}_f$, the regional features $\mathbf{K}^{(M)}$ and $\mathbf{V}^{(M)}$ output by image encoder, and the input word $\mathbf{w}_t$. Generally, the first decoder sub-layer takes $\mathbf{v}_f$ as the query to calculate sparse nonlinear attention with the word embeddings (taken as keys and values) of $\mathbf{w}_t$ to obtain the visual-enhanced word embeddings. Then the output of the first sub-layer will be coupled with the image encoder's output $\mathbf{K}^{(M)}$ and $\mathbf{V}^{(M)}$ as $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ input to another sparse attention block and repeat this process. Note that we employ residual connections around each sub-layers similar to the encoder, followed by layer normalization. The decoder is also composed of a stack of $N = 6$ identical layers.

We train our model parameters θ by minimizing the negative log-likelihood of $\mathbf{P}(T)$ given the image features:

$$\mathcal{L}_{CE} = - \sum_{i=1}^N \log P_{\theta}(\mathbf{t}_i | \mathbf{I}, \mathbf{t}_{i-1}, \dots, \mathbf{t}_1) \quad (5.5)$$

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

Overall objective function Our overall objective integrates the two losses regarding report generation and multi-class classification, which is defined as:

$$\mathcal{L}_{all} = \lambda_{CE} \mathcal{L}_{CE} + \lambda_{MLC} \mathcal{L}_{MLC} \quad (5.6)$$

The hyper-parameters λ_{CE} and λ_{MLC} balance the two losses terms, and their values are given in Section. 5.3.

5.3 Experiments

5.3.1 Data Collection

MIMIC-CXR (Johnson *et al.*, 2019) is the recently released largest dataset to date containing both chest radiographs and free-text reports. It consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center. In our experiment, we adopt MIMIC-CXR’s official split following (Chen *et al.*, 2020b; Liu *et al.*, 2021a) for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test, respectively. We convert all tokens to lowercase characters and remove non-word illegal characters and the words whose occurring frequency is lower than 5, counting to 5412 unique words remaining in the dataset.

5.3.2 Experimental Settings

Evaluation Metrics Following the standard evaluation protocol¹, we utilise the most widely used BLEU scores (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDER (Vedantam *et al.*, 2015) as the metrics to evaluate the quality of the generated text report.

Implementation Details We extract our pre-computed image features by a pre-trained “CLIP-B/16” (Radford *et al.*, 2021)² model, resulting in a regional feature matrix $f \in \mathbb{R}^{768 \times 196}$ (reshaped from $768 \times 14 \times 14$). The number of memory vectors is set to 3. The number of the layers in both the image transformer encoder and the text transformer decoder is set to 6 and the number of heads in multi-head attention is set to 8. The hyper-parameters λ_{CE} , λ_{MLC} are set as 1 and 5, respectively. We train our model using Adam optimizer (Kingma and

¹<https://github.com/tylin/coco-caption>

²<https://github.com/openai/CLIP/>

Table 5.1. Performance comparison on MIMIC-CXR dataset.

Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE	METEOR	CIDEr
Show-Tell	0.308	0.190	0.125	0.088	0.256	0.122	0.096
Att2in	0.314	0.198	0.133	0.095	0.264	0.122	0.106
AdaAtt	0.314	0.198	0.132	0.094	0.267	0.128	0.131
Transformer	0.316	0.199	0.140	0.092	0.267	0.129	0.134
M2Transformer	0.332	0.210	0.142	0.101	0.264	0.134	0.142
R2Gen	0.353	0.218	0.145	0.103	0.277	0.142	0.141
R2GenCMN	0.353	0.218	0.148	0.106	0.278	0.142	0.143
PPKED [†]	0.360	0.224	0.149	0.106	0.284	0.149	0.237
Self-boost	0.359	0.224	0.150	0.109	0.277	0.141	0.270
baseline	0.352	0.215	0.145	0.105	0.264	0.133	0.240
baseline+MSA	0.369	0.231	0.159	0.117	0.280	0.142	0.297
Ours(baseline+MSA+MCGN)	0.373	0.235	0.162	0.120	0.282	0.143	0.299
Ours+RL	0.413	0.266	0.186	0.136	0.298	0.170	0.429

Ba, 2015) on eight NVIDIA TESLA V100 with a mini-batch size of 32 for each GPU. The learning rates is set to be $5e - 5$, and the model is trained in a total of 60 epochs. At the inference stage, we adopt the beam search strategy and set the beam size as 3. In addition, following (Rennie *et al.*, 2017), we also apply reinforcement learning with CIDEr as the reward and further train the model for 20 epochs to boost the performance.

5.3.3 Results and Discussion

Comparison with SOTA Table 5.1 summarizes the performance comparisons between the state-of-the-art methods and our proposed model on the MIMIC-CXR Official test split. Specifically, there are five SOTA image captioning methods in the comparison, including Show-tell (Vinyals *et al.*, 2015), AdaAtt (Lu *et al.*, 2017), Att2in (Anderson *et al.*, 2018), Transformer (Vaswani *et al.*, 2017), and M2transformer (Cornia *et al.*, 2020). Moreover, we also compare with four SOTA medical report generation methods: including R2Gen (Chen *et al.*, 2020b), R2GenCMN (Chen *et al.*, 2021), PPKED (Liu *et al.*, 2021a) and Self-boost (Wang *et al.*, 2021). For PPKED (Liu *et al.*, 2021a), we quote the performance from their paper (marked with [†] in Table. 5.1) since this model does not release the code. For the other methods in comparison, we download the codes

released publicly and re-run them on the MIMIC-CXR dataset with the same experimental setting as ours, so they are strictly comparable.

As shown in Table 5.1, our proposed method is the best performer over almost all evaluation metrics among the comparing methods, even without the reinforcement learning. Specifically, the two very recent medical report generation models, R2Gen (Chen *et al.*, 2020b) and Self-boost (Wang *et al.*, 2021), perform better than other image captioning methods but still lose to ours. R2Gen adopts a memory-augmented transformer decoder, but its image encoder still relies on the CNN model, which may fail to identify fine-grained differences of radiographic images. For the Self-boost model, despite utilizing an image-text matching network to help the image encoder learn fine-grained visual differences, it builds upon CNN and LSTM and does not incorporate semantic information into the report generation process like ours. It is also noted that although performing slightly inferior than PPKED (Liu *et al.*, 2021a) on Meteor and Rouge, our models shows clear advantages over PPKED on Bleu and CIDEr metrics. PPKED is a transformer based model and encodes semantic concepts into an external knowledge graph to guide the report generation. However, this knowledge graph is sparse with only 25 nodes (concepts) and requires extra efforts to construct, while our fine-grained medical concepts can be easily picked up from the reports.

Image	Medical Concepts	Ground Truth	BL	BL+MSXA+MCGN	BL+MSXA+MCGN+RL
	Ground Truth	ap and lateral views of the chest were reviewed. the heart size is top normal. the mediastinal and hilar contours are unremarkable. there is no pleural effusion or pneumothorax. there is no focal consolidation concerning for pneumonia.	the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.	frontal and lateral views of the chest. the heart size is normal. the mediastinal and hilar contours are normal. the lungs are clear. there is no focal consolidation pleural effusion or pneumothorax is seen. no acute osseous abnormalities.
	Predictions	chest, lungs, consolidation, hilar, contours, effusion, osseous, abnormality, pneumothorax, heart			

Figure 5.2. An example report generated by the proposed model. The correctly predicted medical concepts and the key information in the report are marked red.

Ablation Study Table. 5.1(lower part) shows the results of the ablation study to single out the contributions of each component of our model. We remove MCGN from our model and replace MSA with conventional self-attention (Vaswani *et al.*, 2017) as the baseline to verify the performance improvements brought by the proposed attention and our concepts generation network. In table 5.1, there are three components: MSA, MCGN, and RL, representing Memory-augmented Sparse Attention, Medical Concepts Generation Network, and Reinforcement Learning, respectively. The symbols “+” or “-” indicate the inclusion or exclusion of the following component. The benefit of using MSA block can be well reflected by the improvement from “baseline” to “baseline+MSA”. As shown, the performance can be further boosted by additionally introducing medical concepts (“baseline+MSA+MCGN”). Moreover, as mentioned, even removing reinforcement learning, our model still significantly outperforms the comparing methods in Table 5.1. Our proposed model produces more accurate and descriptive findings compared with baseline, as the qualitative results show in Figure. 5.2.

5.4 Conclusions

We propose a medical report generation model utilizing sparse nonlinear attention in the transformer-structured encoder-decoder paradigm. Compared with prior arts, our proposed model considers higher-order interactions across the input feature vectors and thus can better cater for the fine-grained radiographic images. Our proposed Medical Concepts Generation Network introduces rich semantic concepts and encodes them as semantic information to benefit the radiographic report generation task. Extensive experimental results demonstrate that our approach surpasses the state-of-the-art with a large margin.

Automated Radiographic Report Generation Purely On Transformer: A Multi-criteria Supervised Approach

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

6.1 Introduction

Medical report generation roots in a more general research topic of image captioning. A comprehensive review is given in Section 2.1. Despite their achievements, it is known that current image captioning methods usually generate overly rigid reports which do not work well on discriminating similar images and tend to repeat the phrases that frequently appear in the training set (Wu *et al.*, 2020). This makes them difficult to be effectively applied to radiographic report generation due to two challenges. First (Challenge I), medical images are very similar to each other and clinic importance is often reflected via fine-grained visual differences. Second (Challenge II), the disease-related words may be submerged by many similar sentences describing the common content of the images. This could cause the abnormal to be misinterpreted as the normal in the worst case. Existing medical report generation methods have made extra efforts, such as training additional classifiers to predict disease labels or medical tags (Jing *et al.*, 2018; Yin *et al.*, 2019) or utilising knowledge graph as prior (Zhang *et al.*, 2020), to improve the report quality. In addition, since medical reports are long narratives of multiple sentences, efforts have also been spent in developing models, such as hierarchical LSTM (Krause *et al.*, 2017; Jing *et al.*, 2018), for long paragraph generation.

To advance radiographic report generation, we propose a multi-criteria supervised transformer in this chapter. The whole model is built purely upon transformer without the need of CNN. Different from common image captioners using CNN to extract visual features, our transformer based visual encoder circumvents the limited receptive field of CNNs, so that long-range visual dependency could be effectively learned. As for our text decoder, a memory-augmented transformer replaces the commonly used hierarchical LSTM for long paragraph generation, which could fully explore word relationship in long range. Moreover, to respond to the two challenges mentioned above, we bring forward a

multi-criteria supervision framework to better train our transformer for report generation. Specifically, for Challenge I, an objective of image-text matching is tactically proposed to force strongly correlated visual and text features to be learned, so that the generated reports could discriminate similar images. For Challenge II, a multi-label classification objective for medical tag prediction is employed to regularise report generation, as well as a proposed term weighting scheme that adjusts the importance of words for training. Our contributions are summarized as follows.

First, to the best of our knowledge, this is the first work that generates medical reports purely based on a transformer model. Transformer, as arguably the next generation scalable vision model, has made its first step in completely replacing CNN for image classification (Dosovitskiy *et al.*, 2021). Our work demonstrates the possibility for a pure transformer model (under the proposed multi-criteria supervision) to achieve the state-of-the-art performance in medical report generation without using CNN, which has considerable referential importance to successive works. Compared with almost all existing medical report generation methods (Jing *et al.*, 2018; Yin *et al.*, 2019; Yuan *et al.*, 2019; Xue *et al.*, 2018; Xue and Huang, 2019) that rely on a pretrained ResNet (He *et al.*, 2016) or DenseNet (Huang *et al.*, 2017) to extract image-level features for visual representation, our pure-transformer based approach extracts regional features without the need of generating region proposals, and explicitly considers the relationship among image regions for better visual representation. Please note that, although some CNN-based image captioning methods (Anderson *et al.*, 2018; Cornia *et al.*, 2020; Zhou *et al.*, 2020; Yao *et al.*, 2018) use pretrained Fast-RCNN model (Ren *et al.*, 2017b) to detect object regions for regional features, they have to step back to image-level features for medical report generation since such datasets rarely have ground-truth object bounding-boxes to allow fine-tuning Faster-RCNN, while the latter is pretrained on natural images that do not share common object classes with medical images. Building upon the

pure transformer architecture, our method could avoid this problem by circumventing the generation of region proposals when exploring regional features and relationships.

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

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

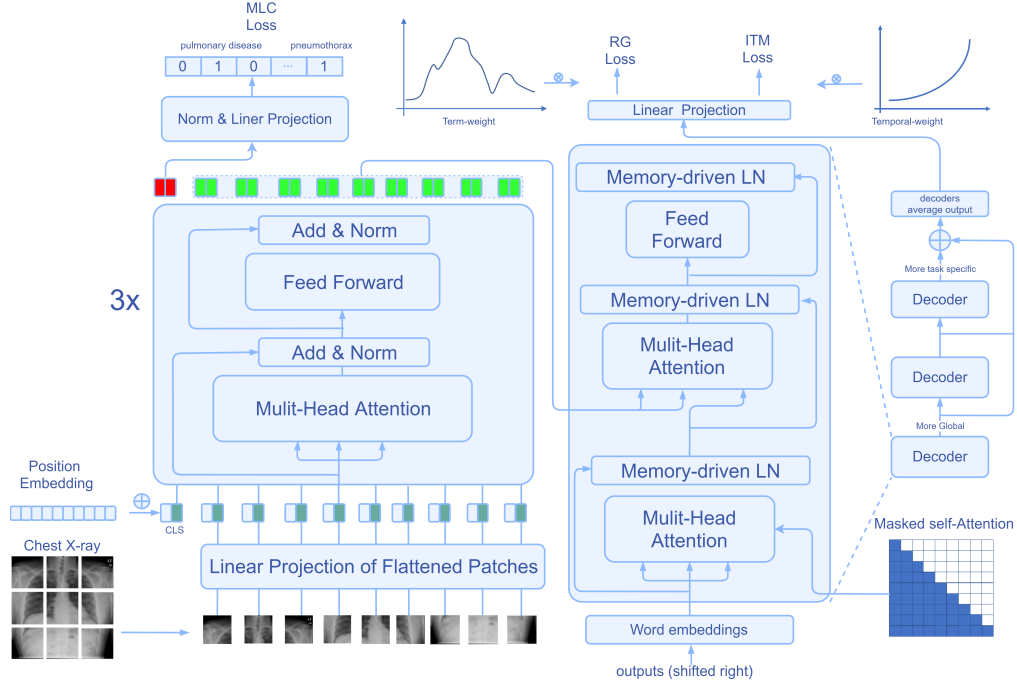


Figure 6.1. An overview of the proposed framework which comprises of a vision transformer encoder and a memory augmented text transformer decoder. The whole model is supervised by three criteria: 1) a term weighting loss for the main task of report generation (RG loss), 2) an image-text matching loss with temporal-weighting (ITM loss), which facilitates report generation by enforcing the model to learn strongly correlated visual and textual features, and 3) a multi-label classification loss (MLC loss) that regularises the report generation by ensuring the model to make correct diagnostic prediction.

6.2 Method

Figure 6.1 shows the overview of our proposed model. It follows the transformer’s encoder-decoder architecture and consists of a vision transformer encoder and a text transformer decoder. The vision encoder splits an image into patches as the input and uses a pure transformer to extract visual features. Its output is fed into both the text decoder to produce a report and the multi-label classification (MLC) network for diagnostic prediction. The text decoder is built upon a memory-augmented transformer which applies a memory-driven layer normalization module (Chen *et al.*, 2020b) to explore and preserve the relation between the words in a report. The generated reports are evaluated by the report

generation (RG) loss via a term-weighting scheme. In addition to the main RG loss and the auxiliary MLC loss, another auxiliary image-text matching (ITM) loss is jointly assessed for training. It evaluates whether a given image-report pair is a true pair, enforcing strongly correlated visual and textual features to be learned for this case. The ITM task shares the same transformer structure as the RG task, with a temporal-weighted matching loss carefully designed by scrutinising transformer’s masked self-attention mechanism. Our model components are elaborated as follows.

6.2.1 Structure of the Proposed Model

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

These image patches are further processed by a linear projection layer and their embedding can be collectively expressed as $\mathbf{E}_p = \mathbf{I}_p \mathbf{W}_p^\top$, where $\mathbf{W}_p \in \mathbb{R}^{D \times P^2 C}$ and therefore $\mathbf{E}_p$ is in the size of $N_p \times D$. Furthermore, at the head of patch embedding, we concatenate a special $[CLS]$ token $\mathbf{E}_{cls}$ like as in Bert (Devlin *et al.*, 2019). It is a learnable one-dimensional embedding of the length D and its output from the transformer encoder (shown as the red block in Figure 6.1) serves as the representation of the entire input image for classification. The concatenated patch embedding $[\mathbf{E}_{cls}; \mathbf{E}_p]$ is then added with a standard learnable 1D position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N_p+1) \times D}$ that carries the positional information,

forming the input $\mathbf{V}_0$ to the transformer encoder:

$$\mathbf{Z}_0 = [\mathbf{E}_{cls}; \mathbf{E}_p] + \mathbf{E}_{pos}, \quad (6.1)$$

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

As shown in Figure 6.1, our vision transformer (Vaswani *et al.*, 2017) encoder consists of three encoder layers (as indicated by “ $3\times$ ” to the left of Figure 6.1) and each layer contains several key components: multi-head self-attention (MSA), Feed Forward (FF), and Layer normalization (LN). The fundamental operation of MSA relies on the concept of self-attention, which is built upon three vectors: queries $\mathbf{q}$, keys $\mathbf{k}$ and values $\mathbf{v}$. They are obtained by mapping the input $\mathbf{z}_0^i$ into a metric space through a linear projection: $\mathbf{q}_i = \mathbf{W}_q \mathbf{z}_0^i$, $\mathbf{k}_i = \mathbf{W}_k \mathbf{z}_0^i$, and $\mathbf{v}_i = \mathbf{W}_v \mathbf{z}_0^i$, where matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable parameters. The output of the MSA operation on $\mathbf{z}_0^i$ is $\mathbf{a}_i = \text{softmax}\left(\frac{\mathbf{q}_i^\top \mathbf{k}}{\sqrt{d_k}}\right) \mathbf{v}_i$, where $i \in [1, N_p + 1]$ and $\mathbf{k} = [\mathbf{k}_1, \dots, \mathbf{k}_{N_p+1}]$, which can be considered as a weighted sum of the value vector while the weight is computed using a softmax activation function over the inner product of the query with the corresponding keys. All the computations can be performed in parallel and expressed as

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}. \quad (6.2)$$

The output of our transformer encoder is as follows:

$$\begin{aligned} \hat{\mathbf{Z}}_l &= \text{MSA}(\text{LN}(\mathbf{Z}_{l-1}) + \mathbf{Z}_{l-1}), l = 1, \dots, L_e \\ \mathbf{Z}_l &= \text{FF}(\text{LN}(\hat{\mathbf{Z}}_{l-1})) + \hat{\mathbf{Z}}_{l-1}, l = 1, \dots, L_e \end{aligned} \quad (6.3)$$

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

of the input image while the $\mathbf{Z}_{L_e}^{pat}$ is fed into the transformer decoder for report generation.

Text Transformer Decoder The input of our text transformer decoder includes the vision encoder’s output $\mathbf{Z}_{L_e}^{pat}$ and the right-shifted text sequence embedding $\mathbf{T}_0$, which is the sum of the word embedding and the positional embedding, as done in common practice. As shown in Figure 6.1, in addition to the common MSA and FF, we employ a memory-driven conditional layer normalization (MLN) (Chen *et al.*, 2020b). Compared with the traditional layer normalization, MLN incorporates the relational memory module proposed in (Chen *et al.*, 2020b) by feeding its output m to γ and β . Consequently, this design can take the benefit from the memory while preventing it from influencing too many parameters of Transformer so that some core information for generation is not affected. The expression of the memory-driven conditional layer normalization is shown as the follows.

$$\text{MLN} = \hat{\gamma} \odot \frac{r - \mu}{\sigma} + \hat{\beta} \quad (6.4)$$

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

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

$$\begin{aligned} \hat{\mathbf{T}}_l &= \text{MSA}(\text{MLN}(\mathbf{T}_{l-1})) + \mathbf{T}_{l-1} \\ \hat{\mathbf{T}}\mathbf{Z}_l &= \text{MSA}(\text{MLN}(\mathbf{Z}_{L_e}^{pat}, \mathbf{Z}_{L_e}^{pat}, \hat{\mathbf{T}}_l)) + \hat{\mathbf{T}}_l \\ \mathbf{T}\mathbf{Z}_l &= \text{FF}(\text{MLN}(\hat{\mathbf{T}}\mathbf{Z}_l)) + \hat{\mathbf{T}}\mathbf{Z}_l \end{aligned} \quad (6.5)$$

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

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

6.2.2 Multi-criteria Objectives

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

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

$$\text{TF}_{sk} = \frac{n_{sk}}{\sum_{s=1}^{N_S} n_{sk}}; \quad \text{IDF}_s = \log \frac{N_K}{|\{k | \mathbf{t}_s \in \mathbf{T}_k\}|}. \quad (6.6)$$

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

$$tw_{sk} = \text{TF}_{sk} \cdot \text{IDF}_s. \quad (6.7)$$

We pre-compute the term weight of each word in a report as $\{tw_i\}_{i=1}^N$, where N denotes the total number of words in the report. By incorporating them into

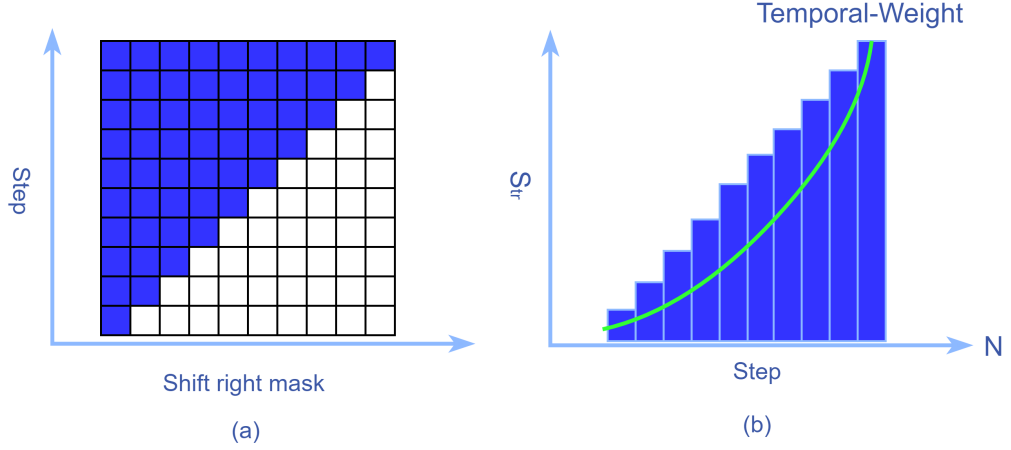


Figure 6.2. (a) The shift-right mask: blue indicates “1” and white indicates “0”. (b) x -axis is the masking step, while y -axis is the ratio of text information (i.e., S_{tr}) in the similarity score. Due to the shift-right masked self-attention, this ratio increases with steps. Accordingly, we assign higher weight (green curve) to the similarity score of later steps to compute image-text matching loss.

the cross-entropy loss commonly used in report generation, we produce a term-weighted report generation loss as follows.

$$\mathcal{L}_{RG}^{tw} = -\frac{\sum_{i=1}^N tw_i \cdot \log P(\mathbf{t}_i | \mathbf{I}, \mathbf{t}_{i-1}, \dots, \mathbf{t}_1)}{\sum_{i=1}^N tw_i}. \quad (6.8)$$

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

Temporal-weighted ITM loss Some visual language tasks such as image-text matching explicitly learn image-text alignment. Compared with image captioning, those tasks do not have to consider sentence grammar and have more objective training losses (Zhou *et al.*, 2020). In this chapter, we leverage the joint training of image-text matching and report generation to learn strongly-correlated visual and textual features to facilitate report generation.

Image-text matching has to be tactically incorporated into our pure transformer-based report generation framework. As discussed in Section 2.4, the sole image-text matching task could be easily embedded into transformer structure as

this only touches the transformer’s encoder and both the visual features and the text embedding are fed into the encoders concurrently. However, the situation becomes more complicated in our case as the image-text matching task is now entangled with the report generation task which contains both the encoder and decoder of the transformer. In this case, we propose to use the encoder for visual feature extraction and incorporate text embedding through the decoder. However, such incorporation is not straightforward since the report generation task and the image-text matching task share the same network model but have different ways of using the image and text features. For report generation, the transformer decoder takes the visual features and the embedding of the previous word as the input to generate the next word. This is implemented by covering the word embedding of the input sentence with a shift-right mask in order to generate the predicted word step by step and meanwhile, compute in parallel, which is known as masked self-attention (Figure 6.2 (a)). However, this shift-right mask will cause an imbalance issue between image and text information during their feature fusion for image-text matching, since in the early steps of the masked self-attention (corresponding to the bottom rows of the mask matrix) a large portion of the sentence is masked off so that only a small portion of text information is used to match the whole image. The text information will gradually increase when more and more words are unmasked (corresponding to the top rows of the mask matrix) and finally the whole sentence is taken into account.

Scrutinising the masked self-attention mechanism, we design a temporal-weighted loss to compensate the imbalance of image and text information. The weighting function is a monotonically increasing one, giving higher weight to the later fused feature since it contains more textual content. It is formally expressed as follows.

Let TE and TD be the embedding functions of the transformer’s encoder and decoder. Let $\mathbf{T}_0 = \{t_i\}_{i=1}^{N_r}$ be the shifted right input sequence of a report, where N_r

is the sequence length. The decoder output is denoted as $\mathbf{F} = \text{TD}(\text{TE}(\mathbf{Z}_0), \mathbf{T}_0)$, where $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^{N_r}$ can be considered as the shift-righted fusion features. Specifically, $\mathbf{f}_1$ is the fusion feature coming from the visual feature $\mathbf{Z}_0$ and the first word t_1 , while $\mathbf{f}_{N_r}$ is the fusion feature from $\mathbf{Z}_0$ and the whole input sentence $\mathbf{T}_0$.

A linear projection layer $\mathbf{S} = \mathbf{W}_f \mathbf{F}$ is further applied to map the fusion feature $\mathbf{F} \in \mathbb{R}^{N_r \times d_k}$ from d_k dimensions to one dimension. We interpret $\mathbf{S} = \{s_i\}_{i=1}^{N_r}$ as the similarity scores of the visual feature $\mathbf{Z}_0$ with the shift-right word tokens. As shown in Figure 6.2 (b), with the increase of the step i , higher weight is assigned to the similarity score s_i . The overall similarity score $S(\mathbf{I}, \mathbf{T})$ between the input image $\mathbf{I}$ and the input text sequence $\mathbf{T}$ is defined as

$$S(\mathbf{I}, \mathbf{T}) = \frac{\sum_{i=1}^{N_r} \lambda_{ITM}^i \cdot s_i}{\sum_{i=1}^{N_r} \lambda_{ITM}^i}, \quad (6.9)$$

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

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

$$\begin{aligned} \mathcal{L}_{ITM} = & [\max(0, S(\mathbf{I}, \mathbf{T}_{pos}) - S(\mathbf{I}, \mathbf{T}_{neg}) + \epsilon)] + \\ & \mu_{ITM} \cdot [S(\mathbf{I}, \mathbf{T}_{neg})^2 + S(\mathbf{I}, \mathbf{T}_{pos})^2], \end{aligned} \quad (6.10)$$

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

Multi-label classification loss. Upon the output of the transformer encoder, we introduce a multi-label classification (MLC) task to predict the medical tags of the input image to regularise report generation. Specifically, we only utilise the embedding of the $[CLS]$ token as the global representation of the image for this

task, denoted as $V_L^{cls} \in \mathbb{R}^{b \times D}$, where b is the number of images in a mini-batch and D is the dimensions of the $[CLS]$ embedding. A linear projection maps V_L^{cls} from D to K dimensions, where K is the number of multi-label classes. The MLC loss is defined as

$$\begin{aligned} \mathcal{L}_{MLC} = & -\frac{1}{K} \cdot \sum_i y_i \cdot \log((1 + \exp(-x_i))^{-1}) \\ & + (1 - y_i) \cdot \log\left(\frac{\exp(-x_i)}{1 + \exp(-x_i)}\right), \end{aligned} \quad (6.11)$$

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

Overall objective function Our overall objective integrates the three losses regarding report generation, image-text matching, and classification, which is defined as

$$\mathcal{L}_{all} = \lambda_{RG} \mathcal{L}_{RG}^{tw} + \lambda_{ITM} \mathcal{L}_{ITM} + \lambda_{MLC} \mathcal{L}_{MLC}. \quad (6.12)$$

The hyper-parameters λ_{RG} , λ_{ITM} and λ_{MLC} balance the three loss terms, and their values are given in Section. 6.3.

6.3 Experiments

6.3.1 Datasets

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

IU-Xray As the most widely used benchmark for medical report generation task, Indiana University Chest X-ray Collection (IU-Xray) (Demner-Fushman *et al.*, 2016) contains 3955 fully de-identified radiology reports, each with a

frontal and/or lateral chest X-ray images, counting to 7470 chest X-ray images in total. Each report is composed of sections such as “Impression”, “Findings” and “Tags”, where “Impression” is a single sentence description of the image, “Findings” is a long paragraph with detailed description of the evidence and “Tags” are the keywords which are identified using the Medical Text Indexer ¹. In this work, by considering only the reports with two complete image views and the complete sections of “Findings” and “Impression”, we work on a smaller dataset with 3195 reports and 6390 images. All words in “Findings” and “Impression” are tokenised into 2076 unique words, with two special tokens, $\langle start \rangle$ and $\langle end \rangle$, indicating the start and the end of a sentence. Following (Jing *et al.*, 2018; Xue *et al.*, 2018), 10% of reports are randomly picked to form the test set. All evaluations are done on the test set.

MIMIC-CXR The recently released MIMIC-CXR (Johnson *et al.*, 2019) is the largest public dataset containing both the chest radiographs and free-text reports. In total, it consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016 (Pavlopoulos *et al.*, 2022). In our experiment, we adopt MIMIC-CXR’s official split following the work (Chen *et al.*, 2020b) for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test, respectively. We convert all tokens to lowercase characters and remove non-word illegal characters and the words whose occurring frequency lower than 10, counting to 4030 unique words remaining in the dataset. In addition, the structured labels provided by this dataset, including “No Finding” and other 13 descriptive labels, are used as the 14 tags to train our MLC task. Each label contains one of the four values: 1, 0, -1 , or “missing”, representing “positively mentioned”, “negatively mentioned”, “mentioned with uncertainty” and “no mention”, respectively. We only keep the labels with the values 1 and 0 when training the MLC task.

¹<https://ii.nlm.nih.gov/MTI/>

6.3.2 Experimental Settings

Evaluation Metrics Following the standard evaluation protocol², we utilise the most widely used BLEU scores (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004) and CIDEr (Vedantam *et al.*, 2015) as the metrics to evaluate the quality of the generated text report.

Implementation Details We use paired images as the input for IU-Xray and concatenate the features of the frontal and lateral views, while use single image input for MIMIC-CXR to ensure consistency with the experiment settings in (Chen *et al.*, 2020b). We also adopted the same preprocessing to the reports as paper (Chen *et al.*, 2020b), removed irregular characters and uniformly replaced the connectors between sentences in the report with “. ”. For vision transformer encoder, we resize the input image to 448×448 and set the patch size to 32. The number of the layers in both the vision transformer encoder and the text transformer decoder is set to 3 and the number of heads in multi-head attention is set to 8. The hyper-parameters ϵ and μ_{ITM}^{Reg} in Eqn. 6.10 are set to 0.05 and 0.025, while λ_{RG} , λ_{MLC} , and λ_{ITM} in Eqn. 6.12 are set as 1, 2, and 5, respectively. We train our model using Adam optimizer (Kingma and Ba, 2015) with mini-batch size of 16. The learning rates are set to be $5e - 5$ and $1e - 4$, respectively, for the vision encoder and the rest parameters. The model is trained in a total of 30 epochs. These settings are equally applied to both datasets. For the model’s efficiency, in the training phase, we use two NVIDIA 2080Ti GPUs and trained our model for about 48 hours on the MIMIC-CXR dataset and 3 hours on the IU-Xray dataset in a total of 30 epochs. In the inference phase, the inference time per batch is about 4.5 seconds, with a batch size of 32 and a maximum report length of 60.

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

Table 6.1. Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from the published literature. Specifically, we quote the results from (Li *et al.*, 2019a) for IU-xray, and from (Chen *et al.*, 2020b) for MIMIC-CXR. For other methods in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method so that they are strictly comparable.

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

6.3.3 Results and Discussion

Comparison with SOTA Methods Table 6.1 shows the performance comparison between our proposed model and existing SOTA methods in image captioning and medical report generation.

On IU-Xray, there are seven SOTA image captioning methods in the comparison, including Show-tell (Vinyals *et al.*, 2015), AdaAtt (Lu *et al.*, 2017), Att2in (Anderson *et al.*, 2018), Topdown (Rennie *et al.*, 2017), Transformer (Vaswani *et al.*, 2017), M2transformer (Cornia *et al.*, 2020), and Grounded (Zhou *et al.*, 2020). The results of these methods are obtained based on our re-running of the publicly released codes under the same experiment setting as ours, so they are strictly comparable. Moreover, we also compare with five SOTA medical report generation methods: including CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et*

et al., 2018), KERP (Li *et al.*, 2019a), MRMA (Xue *et al.*, 2018), and the very recent method R2Gen (Chen *et al.*, 2020b) which incorporated transformer (as the text decoder) for Chest X-ray report generation. For MRMA (Xue *et al.*, 2018) and R2Gen (Chen *et al.*, 2020b), we download the publicly released code and re-run it using the same experiment setting as ours. For the rest three methods CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et al.*, 2018) and KERP (Li *et al.*, 2019a), we have to directly quote their results from (Li *et al.*, 2019a; Chen *et al.*, 2020b) since we could not access their codes. These results were chosen because the performance of the basic Show-Tell method reported in (Li *et al.*, 2019a; Chen *et al.*, 2020b) is close to that obtained in our experiment, implying certain comparison base. However, since the data split is unknown, the quoted results of CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et al.*, 2018) and KERP (Li *et al.*, 2019a) may not be strictly comparable with that of our method, and they were listed just for reference. As shown in Table 6.1 (upper part), on IU-Xray, our proposed method achieves the best performance over almost all evaluation metrics among the comparing methods. It outperforms the second best performer Grounded (Zhou *et al.*, 2020) that also makes use of image-text matching for image captioning over all metrics except BLEU-4. This demonstrates the advantages of our joint training of image-text matching and report generation over the separated manipulation of the two tasks in (Zhou *et al.*, 2020). Meanwhile, it is noted that our proposed model also beats the three transformer based models Transformer (Vaswani *et al.*, 2017), M2transformer (Cornia *et al.*, 2020), and R2Gen (Chen *et al.*, 2020b) in comparison.

On MIMIC-CXR, we compare with R2Gen (Chen *et al.*, 2020b) and another six image captioners as shown in Table 6.1 (lower part). Except M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b), the results of other comparing methods are quoted from (Chen *et al.*, 2020b). Please note that since the official split of MIMIC-CXR is used in all these methods and ours, the results are comparable. Again, our model is overall the best

Table 6.2. Ablation study on key components using IU-Xray. Here “baseline” refers to our pure transformer framework without the three criteria. “MLC”, “ITM”, “TW”, and “TmW” stand for “Multi-Label Classification”, “Image-text Matching”, “Term Weighting”, and “Temporal-Weighting”. BLlue-4 is the BLUE score computed with 4-grams.

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

performer on this largest benchmark. It wins the vanilla Transformer (Devlin *et al.*, 2019), M2transformer (Cornia *et al.*, 2020), and R2Gen (Chen *et al.*, 2020b), all of which still need CNNs in addition to transformer.

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

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

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

Table 6.3. Ablation studies on hyper-parameters of Eqn. 6.9 using IU-Xray

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

we further perform an ablation study to analyze the hyper-parameters in Eqn. 6.9. The Eqn. 6.9 contains two hyper-parameters: ϵ and μ_{ITM} , where ϵ is the margin

of triplet loss and μ_{ITM} is a hyper-parameter for L2-regularization. We fix one hyper-parameter and increase/decrease the other parameter by 2 times, respectively. As shown in Table 6.3, we can achieve a better result when we set ϵ and μ_{ITM} as 0.1 and 0.025 respectively.

Finally, we perform another ablation study based on IU-Xray to investigate the influence of key hyper-parameters λ_{RG} , λ_{ITM} , and λ_{MLC} used to balance our proposed three supervision criteria. We fix $\lambda_{RG} = 1$, and vary the other two parameters as in Table 6.4 for investigation. As can be seen, when these hyper-parameters vary in a reasonably large range, our model can still perform better than most of the methods in comparison in Table 6.1 (upper part).

Qualitative Results Figure 6.3 and Figure 6.4 exhibit the generated reports on the IU-Xray dataset and MIMIC-CXR dataset, respectively. To facilitate comparison, we highlight the key information in ground truth reports in bold. Meanwhile, in the generated reports, sentences are highlighted with different colors: red for incorrect generations (content not existing in the ground-truth reports), blue for partially correct generations (similar meaning but with different expressions), and green for completely correct generations (same expressions). As can be observed, compared with the baseline and the other three competitive methods, the report generated by our full model captures more key information under the supervision of the three criteria, producing more sentences in green, which indicates our generated reports are more consistent with the ground-truth reports.

Table 6.4. Ablation studies on hyper-parameters using IU-Xray. We fix $\lambda_{RG} = 1$ and vary λ_{ITM} and λ_{MLC} for the investigation. The default values are: $\lambda_{RG} = 1$, $\lambda_{ITM} = 5$, and $\lambda_{MLC} = 2$.

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

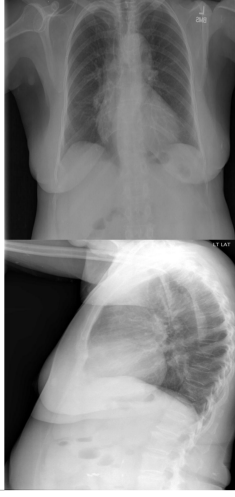
	Ground-truth	M2transformer	Grounded
	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly. Normal pulmonary vascularity. Tortuosity of the descending aorta. No focal infiltrate, pneumothorax or pleural effusion.	Impression: Mild cardiomegaly. Findings: Mild cardiomegaly, the lungs are clear of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

Figure 6.3. Examples of the generated reports on IU-xray dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer (Cornia *et al.*, 2020) ,Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth.

We also add two failure cases as shown in Figure 6.5 for IU-Xray (upper) and MIMIC-CXR (lower) datasets. After analyzing the failure cases, we think one important reason causing the failures might be the long-tailed distribution of sentences/terms in the reports of these datasets. The most commonly used templates in IU-Xray are counted in paper (Jing *et al.*, 2019), including: “The lungs are clear”, “There is no pleural effusion or pneumothorax.”, “No evidence of focal consolidation, pneumothorax, or pleural effusion”, etc. Just like the failure case shown in Figure 6.5, those sentences (marked in red) are more likely be predicted by the model since they appear frequently in the report. Furthermore, for the MIMIC-CXR dataset, we did our own statistics from 222758 training reports to analyze the case shown in Figure 6.5 (lower part). For the generated report, the sentence “pa and lateral views of the chest” appeared 22678 times in the training set, “the lungs are clear” appear 40038 times, the sequence “mild pulmonary edema” appear 5109 times; For the ground-truth report, the

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

Figure 6.4. Examples of the generated reports on MIMIC-CXR dataset. We compare our results with the ground truth report, baseline and the other three most competitive methods: M2transformer (Cornia *et al.*, 2020), Grounded (Zhou *et al.*, 2020) and R2Gen (Chen *et al.*, 2020b). The highlighted sentences with red, blue and green color in the generated reports respectively represent: incorrectly predicted sentences; sentences not exactly consistent with ground-truth; sentences that completely match the ground-truth.

term “mild-to-moderate pulmonary edema” appears 432 times and the sentence “small bilateral pleural effusions have resolved” appears only 19 times in all training reports, while they are missing in the generated report. From the above analysis, we can see that the problem of long-tail distribution does exist and this will be a very valuable research direction in the future.

6.4 Conclusion

This work is the first attempt to build pure transformer based model to accurately generate diagnostic reports from radiographs. The proposed model is effectively trained under three supervision criteria to ensure well aligned visual and textual features, accurate multi-label diagnostic discrimination, and word-importance

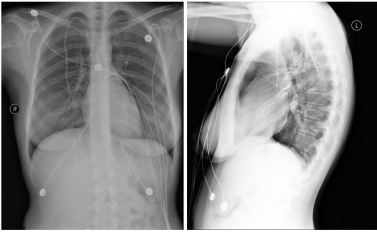
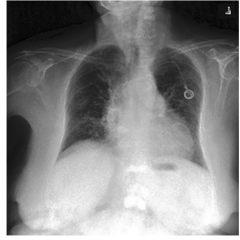
IU-Xray	
	ground-truth report Impression: No acute findings Heart size within normal limits . Findings: stable mediastinal and hilar contours, right chest XXXX tip in the low SVC. Monitoring device artifacts. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema.
	generated report Impression: no acute cardiopulmonary abnormality . Findings: the cardiomeastinal silhouette is normal in size and contour . the lungs are clear . there is no pleural effusion or pneumothorax . there is no focal air space opacity . no pleural effusion or pneumothorax . the osseous structures are intact .
MIMIC-CXR	
	ground-truth report Impression: Improved but not resolved mild-to-moderate pulmonary edema. Findings: there has been improvement in mild-to-moderate pulmonary edema with decreased interstitial markings compared to most recent prior study. small bilateral pleural effusions have resolved. there is no focal consolidation or pneumothorax. heart size is moderately enlarged and stable. a left chest wall port-a-cath terminates in the ra. the patient is status post multiple vertebroplasties
	generated report Impression: Mild pulmonary edema Findings: pa and lateral views of the chest . the lungs are clear . there is mild pulmonary edema . there is no pleural effusion or pneumothorax . the heart size is mildly enlarged . the mediastinal and hilar contours are normal . there is no focal consolidation.

Figure 6.5. The failure cases for IU-xray dataset (upper case) and MIMIC-CXR dataset (lower case). Sentences marked in red indicate frequently occurring sentences in the report

aware report generation. In this way, the generated reports are sensitive to similar images and key words of clinic importance, offering the state-of-the-arts performance on two benchmarks. Although our method achieved promising results in medical report generation, there are some limitations in the current model. First, the example failure cases we provided in Figure 6.5 show the evidence about the problem of long-tail distribution of terms in the training reports. There are some efforts exploring long-tail distribution in the field of natural image classification and object detection, but such research in the natural language processing and medical report generation is still superficial and could be taken as a future research direction. Second, we apply global self-attention for the report, making the calculation relatively expensive. Efficient transformers (Tay et al., 2023) could be explored in the future for medical report generation. Furthermore, as for our future work, the medical text indexer can be utilized to obtain more fine-grained tags for multi-label classification task. Moreover, knowledge

graphs could potentially be built into our transformer-based framework for further improvement with additional information, as knowledge graphs may be beneficial to the generation of medical reports due to their ability to identify and model the relationships between different medical terms. However, the knowledge graphs constructed in the existing works (Zhang *et al.*, 2020; Liu *et al.*, 2021a) are pretty sparse with very minimal information, requiring further improvement in future.

METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens

In clinical scenarios, multi-specialist consultation could significantly benefit the diagnosis, especially for intricate cases. This inspires us to explore a “multi-expert joint diagnosis” mechanism to upgrade the existing “single expert” framework commonly seen in the current literature. To this end, we propose METransformer, a method to realize this idea with a transformer-based backbone. The key design of our method is the introduction of multiple learnable “expert” tokens into both the transformer encoder and decoder. In the encoder, each expert token interacts with both vision tokens and other expert tokens to learn to attend different image regions for image representation. These expert tokens are encouraged to capture complementary information by an orthogonal loss that minimizes their overlap. In the decoder, each attended expert token guides the cross-attention between input words and visual tokens, thus influencing the generated report. A metrics-based expert voting strategy is further developed to generate the final report. By the multi-experts concept, our model enjoys the merits of an ensemble-based approach but through a manner that is computationally more efficient and supports more sophisticated interactions among experts. Experimental results demonstrate the promising performance of our proposed model on two widely used benchmarks. Last but not least, the framework-level innovation makes our work ready to incorporate advances on existing “single-expert” models to further improve its performance.

7.1 Introduction

Interpreting radiology images (e.g., chest X-ray) and writing diagnostic reports are essential operations in clinical practice and normally require a considerable manual workload. Therefore, radiology report generation, which aims to automatically generate a free-text description based on a radiograph, is highly desired to ease the burden of radiologists while maintaining the quality of health care. Recently, substantial progress has been made towards research on automated radiology report generation models (Li *et al.*, 2018; Li *et al.*, 2019a; Zhang *et al.*, 2020; Yin *et al.*, 2019; Chen *et al.*, 2020b; Li *et al.*, 2023b; Wang *et al.*, 2021; Liu *et al.*, 2021a; Chen *et al.*, 2021; Yang *et al.*, 2021b; Wang *et al.*, 2022c; You *et al.*, 2021; Liu *et al.*, 2021b; Wang *et al.*, 2022b). Most existing studies adopt a conventional encoder-decoder architecture following the image captioning paradigm (Vinyals *et al.*, 2015; Xu *et al.*, 2015; Cornia *et al.*, 2020; Zhou *et al.*, 2020; Pan *et al.*, 2020) and resort to optimizing network structure or introducing external or prior information to aid report generation. These methods, in this chapter, are collectively referred to as “single-expert” based diagnostic captioning methods.

However, diagnostic report generation is a very challenging task as disease anomalies usually only occupy a small portion of the whole image and could appear at arbitrary locations. Due to the fine-grained nature of radiology images, it is hard to focus on the correct image regions throughout the report generation procedure despite different attentions developed in recent works (Li *et al.*, 2023b; Liu *et al.*, 2021a). Meanwhile, it is noticed that in clinic scenarios, multi-specialist consultation is especially beneficial for those intricate diagnostic cases that challenge a single specialist for a comprehensive and accurate diagnosis. The above observations led us to think, could we design a model to simulate the multi-specialist consultation scenario? Based on this motivation, we propose a new diagnostic captioning framework, METransformer, to mimic the “multi-expert joint diagnosis” process. Built upon a transformer backbone,

METransformer introduces multiple “expert tokens”, representing multiple experts, into both the transformer encoder and decoder. Each expert token learns to attend distinct image regions and interacts with other expert tokens to capture reliable and complementary visual information and produce a diagnosis report in parallel. The optimal report is selected through an expert voting strategy to produce the final report. Our design is based on the assumption that it would be easier for multiple experts than a single one to capture visual patterns of clinical importance, which is verified by our experimental results.

Specifically, we feed both the expert tokens (learnable embeddings) and the visual tokens (image patches embeddings) into the Expert Transformer encoder which is comprised of a vision transformer (ViT) encoder and a bilinear transformer encoder. In ViT encoder, each expert token interacts not only with the visual tokens but also with the other expert tokens. Further, in the bilinear transformer encoder, to enable each “expert” to capture fine-grained image information, we compute higher-order attention between expert tokens and visual tokens, which has proved to be effective in fine-grained classification tasks (Lin *et al.*, 2015). It is noteworthy that the expert tokens in the encoder are encouraged to learn complementary representations by an orthogonal loss so that they attend differently to different image regions. With these carefully learned expert token embeddings, in the decoder, we take them as a guide to regulate the learning of word embeddings and visual token embedding in the report generation process. This results in M different word and visual token embeddings, thus producing M candidate reports, where M is the number of experts. We further propose a metric-based expert voting strategy to generate the final report from the M candidates.

By utilizing multiple experts, our model, to some extent, is analogous to ensemble-based approaches, while each expert token corresponds to an individual model. While it enjoys the merits of ensemble-based approaches, our model is designed

in a manner that is computationally more efficient and supports more sophisticated interactions among experts. Therefore, it can scale up with only a trivial increase of model parameters and achieves better performance, as demonstrated in our experimental study.

Our main contributions are summarized as follows.

First, we propose a new diagnostic captioning framework, METransformer, which is conceptually “multi-expert joint diagnosis” for radiology report generation, by introducing learnable expert tokens and encouraging them to learn complementary representations using both linear and non-linear attentions.

Second, our model enjoys the benefits of an ensemble approach. Thanks to the carefully designed network structure and the end-to-end training manner, our model can achieve better results than common ensemble approaches while greatly reducing training parameters and improving training efficiency.

Third, our approach shows promising performance on two widely used benchmarks IU-Xray and MIMIC-CXR over multiple state-of-the-art methods. The clinic relevance of the generated reports is also analyzed.

7.2 Method

As shown in Figure.7.1, our encoder comprises a vision transformer (ViT) (Doso-vitskiy *et al.*, 2021) encoder and an expert bilinear transformer encoder. The ViT encoder takes both the expert and the visual tokens as the input and computes the linear attention between every two tokens. The encoded tokens are further sent into the expert bilinear transformer encoder where high-order interactions are computed between the expert tokens and the visual tokens. The enhanced expert tokens are then used to regulate the embeddings of visual and word tokens by an adjust block and sent into the expert decoder to produce expert-specific

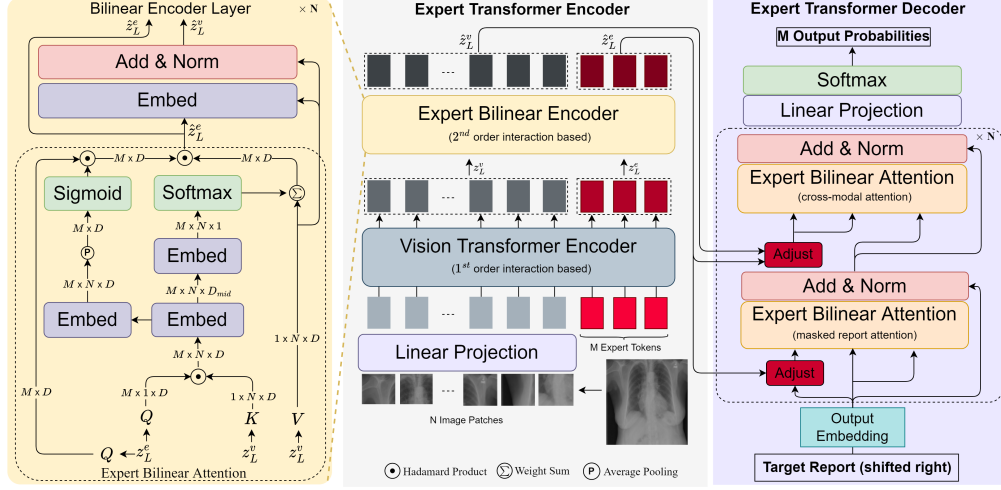


Figure 7.1. An overview of our proposed METransformer, which includes an Expert Transformer Encoder and an Expert Transformer Decoder. The expert transformer encoder captures fine-grained visual patterns by exploring 1st and 2nd order interactions between input visual tokens and M expert tokens. The expert transformer decoder produces M diverse candidate reports guided by each expert token through an Adjust block. Noted that the “Embed” layer includes a linear projection followed by a ReLU activation function.

reports in parallel, and then an expert voting strategy is used to generate the final report.

7.2.1 Multi-expert ViT Encoder

Our ViT encoder adopts vision transformer (Dosovitskiy *et al.*, 2021). In addition to the common input of visual patches/tokens embedding and position embedding, we further introduce expert tokens embedding and segment embedding.

Visual patches Embedding. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, followed (Dosovitskiy *et al.*, 2021), we reshape $\mathbf{x}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \dot{C})}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. We consider $\mathbf{x}_p$ as visual tokens and N is the input sequence length.

Expert tokens Embedding. In addition to N visual tokens, we post-pend M learnable embeddings $\mathbf{x}_e \in \mathbb{R}^{M \times D}$ which have the same dimension as the visual token embeddings and are called expert tokens. We introduce an Orthogonal Loss as in Eqn. 7.6 to enforce orthogonality among the expert token embeddings, encouraging different expert tokens to attend different image regions.

Segment Embedding. We introduce two types of the segment, “[0]” and “[1]”, to separate the input tokens from different sources (Su *et al.*, 2020), i.e., “[0]” for visual tokens and “[1]” for expert tokens. A learned segment embedding $\mathbf{E}_{seg}$ is added to each input token to indicate its segment type.

Position Embedding. A standard 1D learnable position embedding $\mathbf{E}_{pos}$ is added to each input token to indicate its order in the input sequence.

Model Structure. The expert ViT encoder adopts the same structure as the standard Transformer (Vaswani *et al.*, 2017), which consists of alternating layers of multi-headed self-attention (MSA) and Multi-Layer Perceptron blocks (MLP). Layernorm(LN) (Ba *et al.*, 2016) is applied before every block and residual connections are applied after every block. Mathematically:

$$\begin{aligned} \mathbf{z}_0 &= [\mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}; \mathbf{x}_e^1; \mathbf{x}_e^2; \dots; \mathbf{x}_e^M] \\ &\quad + \mathbf{E}_{pos} + \mathbf{E}_{seg}, \\ \hat{\mathbf{z}}_l &= \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \\ \mathbf{z}_l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}_l)) + \hat{\mathbf{z}}_l, \end{aligned} \tag{7.1}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 C) \times D}$ is a learnable matrix parameter to map visual tokens to a fixed dimension D . The subscript $l = 1 \dots L$ and L is the total number of the transformer layers. $\mathbf{E}_{pos} \in \mathbb{R}^{(N+M) \times D}$ is the position embeddings and $\mathbf{E}_{seg} \in \mathbb{R}^{(N+M) \times D}$ is the segment embeddings.

7.2.2 Multi-expert Bilinear Attention Encoder

The output of our expert ViT encoder, $\mathbf{z}_L \in \mathbb{R}^{(N+M) \times D}$, combines the visual token embeddings $\mathbf{z}_L^v = \mathbf{z}_L[:N]$ and the expert token embeddings $\mathbf{z}_L^e = \mathbf{z}_L[N:(N+M)]$. They have attended to each other using multi-head self-attention, which is linear attention. Considering the fine-grained nature of medical images and the effectiveness of high-order attention in fine-grained recognition (Lin *et al.*, 2015), we further enhance both embeddings by expert bilinear attention.

Expert Bilinear Attention. We followed (Pan *et al.*, 2020) to design our expert bilinear attention (EBA) module. As shown in Figure. 7.1 Left, we take the enhanced expert token embeddings $\mathbf{z}_L^e$ as query $\mathbf{Q} \in \mathbb{R}^{M \times 1 \times D_q}$, the enhanced visual tokens embeddings $\mathbf{z}_L^v$ as key $\mathbf{K} \in \mathbb{R}^{1 \times N \times D_k}$ and value $\mathbf{V} \in \mathbb{R}^{1 \times N \times D_v}$, a low-rank bilinear pooling (Kim *et al.*, 2017) if first performed to obtain the joint bilinear query-key $\mathbf{B}_k$ and query-value $\mathbf{B}_v$ by $\mathbf{B}_k = \sigma(\mathbf{W}_k \mathbf{K}) \odot \sigma(\mathbf{W}_q^k \mathbf{Q})$ and $\mathbf{B}_v = \sigma(\mathbf{W}_v \mathbf{V}) \odot \sigma(\mathbf{W}_q^v \mathbf{Q})$, where $\mathbf{W}_k \in \mathbb{R}^{D_B \times D_k}$, $\mathbf{W}_v \in \mathbb{R}^{D_B \times D_v}$, $\mathbf{W}_q^k \in \mathbb{R}^{D_B \times D_q}$ and $\mathbf{W}_q^v \in \mathbb{R}^{D_B \times D_v}$ are learnable parameters, resulting $\mathbf{B}_k \in \mathbb{R}^{M \times N \times D_B}$, $\mathbf{B}_v \in \mathbb{R}^{M \times N \times D_B}$. σ denotes ReLU unit, and $\odot$ represents Hadamard Product. Afterward, we compute attention for input tokens both spatial and channel-wisely. 1) Spatial-wise attention. We use a linear layer to project $\mathbf{B}_k$ into an intermediate representation $\mathbf{B}_{mid} = \sigma(\mathbf{W}_B^k \mathbf{B}_k)$, where $\mathbf{W}_B^k \in \mathbb{R}^{D_B \times D_{mid}}$. Then another linear layer is applied to mapping $\mathbf{B}_{mid} \in \mathbb{R}^{M \times N \times D_{mid}}$ from the dimension D_{mid} to 1 and further followed a softmax function to obtain the spatial-wise attention weight $\alpha_s \in \mathbb{R}^{M \times N \times 1}$. 2) Channel-wise attention. A squeeze-excitation operation (Hu *et al.*, 2018) to $\mathbf{B}_{mid}$ is performed to obtain channel-wise attention $\beta_c = \text{sigmoid}(\mathbf{W}_c \mathbf{B}_{mid}^-)$, where $\mathbf{W}_c \in \mathbb{R}^{D_{mid} \times D_B}$ is learnable parameters and $\mathbf{B}_{mid}^- \in \mathbb{R}^{M \times D_{mid}}$ is the average pooling of $\mathbf{B}_{mid}$. The first layer of EBA is formulated as,

$$\hat{\mathbf{z}}_L^{e(1)} = \text{EBA}(\hat{\mathbf{z}}_L^e, \hat{\mathbf{z}}_L^v) = \beta_c \odot \alpha_s \mathbf{B}_v. \quad (7.2)$$

Bilinear Encoder Layer. Besides EBA, we also use the “Add & Norm” layer with the residual connection as in a standard transformer in our bilinear encoder layer. The n -th layer bilinear encoder is expressed as,

$$\begin{aligned}\hat{\mathbf{z}}_L^{e(n)} &= \text{EBA}(\hat{\mathbf{z}}_L^{e(n-1)}, \hat{\mathbf{z}}_L^{v(n-1)}) \\ \hat{\mathbf{z}}_L^{v(n)} &= \text{LN}(\mathbf{W}_e^n [\hat{\mathbf{z}}_L^{e(n-1)}; \hat{\mathbf{z}}_L^{v(n-1)}] + \hat{\mathbf{z}}_L^{v(n-1)})\end{aligned}\quad (7.3)$$

where $n = 1 \dots N$ and N is the number of bilinear encoder layer, $\text{LN}(\cdot)$ denotes layer normalization (Ba *et al.*, 2016), $\mathbf{W}_e^n \in \mathbb{R}^{(D_q+D_v) \times D_v}$ is learnable parameters and $[\cdot]$ denotes concatenation. We set $\hat{\mathbf{z}}_L^{e(0)} = \mathbf{z}_L^e$ and $\hat{\mathbf{z}}_L^{v(0)} = \mathbf{z}_L^v$ when $n=1$.

7.2.3 Multi-expert Bilinear Attention Decoder

The output of the last expert bilinear encoder, expert token embeddings $\hat{\mathbf{z}}_L^{e(N)} \in \mathbb{R}^{N \times D_B}$ and visual tokens embeddings $\hat{\mathbf{z}}_L^{v(N)} \in \mathbb{R}^{M \times D_B}$, are sent to the decoder for report generation. The bilinear decoder layer comprises an EBA_{mask} layer to compute attention of the shifted-right reports and an EBA_{cross} to compute the cross-modal attention and an adjust block to regulate the word and visual tokens embeddings by expert token embeddings. For convenience, we denote $\hat{\mathbf{z}}_L^{e(N)}$ and $\hat{\mathbf{z}}_L^{v(N)}$ as $\mathbf{f}_e$ and $\mathbf{f}_v$.

Adjust block. To incorporate the expert tokens into the report generation process, we propose an expert adjustment block that allows each expert token embedding to influence the embedding of the words and visual tokens, thereby generating the report associated with that expert token. Since our expert tokens are trained orthogonally, we can generate discrepant reports by the different expert tokens. To regulate visual tokens embeddings $\mathbf{f}_v$ by expert token embeddings $\mathbf{f}_e$, the adjust block is calculated as follows,

$$\hat{\mathbf{f}}_v = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{f}_v) = \sigma(\mathbf{W}_e \mathbf{f}_e) \odot \sigma(\mathbf{W}_v \mathbf{f}_v) \quad (7.4)$$

where $\mathbf{W}_e$ and $\mathbf{W}_v$ are learnable parameters. σ denotes ReLU unit, and $\odot$ represents Hadamard Product.

Bilinear Decoder Layer. We first perform a masked EBA to word embeddings $\mathbf{E}_r$ and then followed another EBA block to compute the cross attention of word embeddings and visual tokens embeddings. We also employ residual connections around each EBA block similar to the encoder, followed by layer normalization. Mathematically, the i -th layer bilinear decoder can be expressed as,

$$\begin{aligned}\mathbf{E}_{mid}^{(i)} &= \text{LN}(\text{EBA}_{mask}(\hat{\mathbf{E}}_r^{(i-1)}, \mathbf{E}_r^{(i-1)}) + \mathbf{E}_r^{(i-1)}) \\ \mathbf{E}_c^{(i)} &= \text{LN}(\text{EBA}_{cross}(\mathbf{E}_{mid}^{(i)}, \hat{\mathbf{f}}_v) + \mathbf{E}_{mid}^{(i)}) \\ \mathbf{E}_r^{(i)} &= \text{LN}(\mathbf{W}_d^i[\mathbf{E}_r^{(i-1)}; \mathbf{E}_c^{(i)}] + \mathbf{E}_r^{(i-1)})\end{aligned}\quad (7.5)$$

where $i = 1 \dots I$ and I represents the total number of decoder layer. $\hat{\mathbf{E}}_r^{(i-1)} = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{E}_r^{(i-1)})$ and $\hat{\mathbf{f}}_v = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{f}_v)$. $\mathbf{W}_d^i \in \mathbb{R}^{(D_r+D_r) \times D_r}$ is learnable parameters and $[\cdot]$ denotes concatenation. Specifically, $\mathbf{E}_r^{(0)} \in \mathbb{R}^{T \times D_r}$ is the original word embeddings where T is the total number of words in the reports and $D_r = D_B$ is the dimension of word embedding, and we extend it to M replicates corresponding to M expert tokens to compute parallelly. The final output of the decoder is $\mathbf{E}_c^I \in \mathbb{R}^{M \times T \times D_r}$, which will be further used to predict word probabilities by a linear projection and the softmax activation function.

7.2.4 Objective Function

Orthogonal Loss. To encourage orthogonality among expert token embeddings, we introduce an orthogonal loss term to the output of expert bilinear encoder $\hat{\mathbf{z}}_L^e$,

$$\mathcal{L}_{OrL}(\hat{\mathbf{z}}_L^e) = \frac{1}{M} \left\| \ell(\hat{\mathbf{z}}_L^e)^\top \ell(\hat{\mathbf{z}}_L^e) - \mathbf{I} \right\|^2 \quad (7.6)$$

where $\ell(\cdot)$ denotes L_2 normalization and $\mathbf{I}$ represents the identity matrix with dimension M .

Report Generation Loss. We train our model parameters θ by minimizing the negative log-likelihood of $\mathbf{P}(t)$ given the image features:

$$\mathcal{L}_{CE} = -\frac{1}{M} \sum_{m=1}^M \sum_{i=1}^T \log P_{\theta}(\mathbf{t}_i^{(m)} | \mathbf{I}, \mathbf{t}_{i-1}^{(m)}, \dots, \mathbf{t}_1^{(m)}) \quad (7.7)$$

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

Our overall objective function is: $\mathcal{L}_{all} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{OrL}$. The hyper-parameter λ simply balances the two loss terms, and its value is given in Section 7.3.

7.2.5 Expert Voting strategy

For the M diagnostic reports $R = [r_1, r_2, \dots, r_M]$ produced by M experts, we design a metric-based expert voting strategy to select the optimal one, where the “metric” means the conventional natural language generation (NLG) metrics, such as BLEU-4 and CIDEr (we use CIDEr in our paper). The voting score $\mathbf{S}_i$ for i -th expert’s report can be calculated by the following equation:

$$\mathbf{S}_i = \sum_{j=1, j \neq i}^M \text{CIDEr}(\mathbf{r}_i, \mathbf{r}_j) \quad (7.8)$$

where $\text{CIDEr}(\cdot)$ denotes the function for computing CIDEr with $\mathbf{r}_i$ as candidate and $\mathbf{r}_j$ as reference. In this way, each expert’s report can get a vote score, indicating the degree of consistency of the report with that of other experts. The diagnostic report with the highest score is the winner of the voting. As demonstrated in Table. 7.3, our voting strategy is more effective than the commonly used ensemble/fusion methods (Pan *et al.*, 2020). The possible reasons are that i) our method utilizes NLG metrics as reference scores, thereby the voted results are directly related to the final evaluation metrics; ii) since we ultimately select a single result with the highest score, we can produce a more coherent report than fusing multiple results at the word level.

7.3 Experiments

7.3.1 Datasets

In this experiment, two datasets are used for the performance evaluation, i.e., a widely-used benchmark IU-Xray (Demner-Fushman *et al.*, 2016) and the currently largest dataset MIMIC-CXR (Johnson *et al.*, 2019) for medical report generation.

IU-Xray Indiana University Chest X-ray Collection (IU-Xray) (Demner-Fushman *et al.*, 2016) is the most widely used publicly accessible dataset in medical report generation tasks. It contains 3955 fully de-identified radiology reports, each of which is associated with frontal and/or lateral chest X-ray images, and 7470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings, Indication, etc. In this work, we adopt the same data set partitioning as (Chen *et al.*, 2020b) for a fair comparison, with a train/test/val set by 7:1:2 of the entire dataset. All evaluations are done on the test set.

MIMIC-CXR The recently released MIMIC-CXR (Johnson *et al.*, 2019) is the largest public dataset containing both chest radiographs and free-text reports. In total, it consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016. In our experiment, we adopt MIMIC-CXR’s official split following the work (Chen *et al.*, 2020b) for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test.

7.3.2 Experimental Settings

Evaluation Metrics Following the standard evaluation protocol¹, we utilize the most widely used BLEU-4 (Papineni *et al.*, 2002), METEOR (Banerjee and

¹<https://github.com/tylin/coco-caption>

Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam *et al.*, 2015) as the metrics to evaluate the quality of the generated diagnostic reports. To measure the accuracy of descriptions for clinical abnormalities, we follow (Chen *et al.*, 2020b; Chen *et al.*, 2021; Liu *et al.*, 2021b) and further report clinical efficacy metrics. For this purpose, the CheXpert (Irvin *et al.*, 2019) is applied to labeling the generated reports and the results are compared with ground truths in 14 different categories related to thoracic diseases and support devices. We use precision, recall, and F1 to evaluate model performance for clinical efficacy metrics.

Methods	Precision	Recall	F1
Show-Tell (Vinyals <i>et al.</i> , 2015)	0.249	0.203	0.204
Att2in (Xu <i>et al.</i> , 2015)	0.268	0.186	0.181
AdaAtt (Lu <i>et al.</i> , 2017)	0.322	0.239	0.249
Transformer (Vaswani <i>et al.</i> , 2017)	0.331	0.224	0.228
R2Gen (Chen <i>et al.</i> , 2020b)	0.333	0.273	0.276
R2GenCMN (Chen <i>et al.</i> , 2021)	0.334	0.275	0.278
Ours(METransformer)	0.364	0.309	0.311

Table 7.1. Comparison of clinical efficacy metrics on the test set of the MIMIC-CXR dataset for measuring the accuracy of the description of clinical abnormalities.

Implementation Details For IU-Xray, we use image pairs for report generation as (Chen *et al.*, 2020b). For both datasets, we use the “bert-base-uncased” model’s tokenizer in huggingface transformer (Wolf *et al.*, 2020) to tokenize all words in the reports. We utilize a pre-trained vision transformer with a patch size of 32 to initialize our Expert ViT encoder. The number of layers of the expert bilinear encoder and decoder is set as (2, 4) for MIMIC-CXR and (2,2) for IU-Xray to reduce the potential overfitting on IU-Xray due to its relatively small data size. The dimensions of the bilinear query-key representation and the transformed bilinear feature (D_B and D_{mid}) in the expert bilinear attention block is set as 768 and 384, respectively. The hyper-parameter λ is set to 2. Our model is trained using Radam optimizer (Liu *et al.*, 2020) with a mini-batch size of 16. The learning rate is set to be 0.0001 and the model is trained for a total

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
IU-Xray	Show-Tell	0.243	0.130	0.108	0.078	0.307	0.157	0.197
	Att2in	0.248	0.134	0.116	0.091	0.309	0.162	0.215
	AdaAtt	0.284	0.207	0.150	0.126	0.311	0.165	0.268
	Transformer	0.372	0.251	0.147	0.136	0.317	0.168	0.310
	M2transformer	0.402	0.284	0.168	0.143	0.328	0.170	0.332
	R2Gen [†]	0.470	0.304	0.219	0.165	0.371	0.187	-
	R2GenCMN [†]	0.475	0.309	0.222	0.170	0.375	0.191	-
	MSAT	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	Ours(METransformer)	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	Results below are not strictly comparable due to different data partition. For reference only.							
	CoAtt [†]	0.455	0.288	0.205	0.154	0.369	-	0.277
	HGRG-Agent [†]	0.438	0.298	0.208	0.151	0.322	-	0.343
	KERP [†]	0.482	0.325	0.226	0.162	0.339	-	0.280
	PPKED [†]	0.483	0.315	0.224	0.168	0.376	0.187	0.351
	GSK [†]	0.496	0.327	0.238	0.178	0.381	-	0.382
MIMIC-CXR	Show-Tell	0.308	0.190	0.125	0.088	0.256	0.122	0.096
	Att2in	0.314	0.198	0.133	0.095	0.264	0.122	0.106
	AdaAtt	0.314	0.198	0.132	0.094	0.267	0.128	0.131
	Transformer	0.316	0.199	0.140	0.092	0.267	0.129	0.134
	M2Transformer	0.332	0.210	0.142	0.101	0.264	0.134	0.142
	R2Gen [†]	0.353	0.218	0.145	0.103	0.277	0.142	-
	R2GenCMN [†]	0.353	0.218	0.148	0.106	0.278	0.142	-
	PPKED [†]	0.36	0.224	0.149	0.106	0.284	0.149	0.237
	GSK [†]	0.363	0.228	0.156	0.115	0.284	-	0.203
	MSAT [†]	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	Ours(METransformer)	0.386	0.250	0.169	0.124	0.291	0.152	0.362

Table 7.2. Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). [†] indicates the results are quoted from the published literature: the results of CoAtt (Jing *et al.*, 2018) and HGRG-Agent (Li *et al.*, 2018) on IU-Xray are quoted from (Li *et al.*, 2019a) while the other results are quoted from their respective papers. For the methods without [†], their results are obtained by re-running the publicly released codebase (Li *et al.*, 2021) on these two datasets using the same training-test partition as our method.

of 20 epochs. We implement our model using Pytorch (Paszke *et al.*, 2019) and Pytorch-lightning library² with two NVIDIA GeForce RTX 3090 GPU cards.

7.3.3 Main Results

Two types of metrics are used in our evaluation: the conventional natural language generation (NLG) metrics and the clinical efficacy (CE) metrics. The results are reported in Table 7.2 and Table. 7.1³, respectively.

²<https://github.com/Lightning-AI/lightning>

³It is noted that clinical efficacy metrics only apply to MIMIC-CXR because the labeling schema of CheXpert is designed for MIMIC-CXR.

Specifically, we compare METransformer with 5 state-of-the-art (SOTA) image captioning methods, including Show-tell (Vinyals *et al.*, 2015), AdaAtt (Lu *et al.*, 2017), Att2in (Anderson *et al.*, 2018), Transformer (Devlin *et al.*, 2019) and M2transformer (Cornia *et al.*, 2020). For these methods, we use a publicly released codebase (Li *et al.*, 2021) and re-run them on both datasets with the same experimental setting as ours. Moreover, eight SOTA medical report generation models are involved in the comparison, including CoAtt (Jing *et al.*, 2018), HGRN-Agent (Li *et al.*, 2018), R2Gen (Chen *et al.*, 2020b), R2GenCMN (Chen *et al.*, 2021), PPKED (Liu *et al.*, 2021a), GSK (Yang *et al.*, 2021b) and MSAT (Wang *et al.*, 2022c). It is noteworthy that except R2Gen (Chen *et al.*, 2020b) and R2GenCMN (Chen *et al.*, 2021), these methods do not have their source code released. CoAtt (Jing *et al.*, 2018) and HGRN-Agent (Li *et al.*, 2018) only report results on IU-Xray in their original paper. For the others, we cite the results from their respective papers. Please note that since IU-Xray does not provide an official training-test partition, the cited performances of these methods on IU-Xray (except R2Gen (Chen *et al.*, 2020b) and R2GenCMN (Chen *et al.*, 2021) that use the same partition as ours) are not strictly comparable, and they are provided here only for reference. Differently, on MIMIC-CXR, since all these models follow the MIMIC-CXR official training-test partition, their cited performances are comparable.

As shown in Table 7.2, on both datasets, our METransformer consistently outperforms those “single-expert” based models, including attention-based baselines (Att2in (Xu *et al.*, 2015), AdaAtt (Lu *et al.*, 2017)), memory-augmented methods (R2Gen (Chen *et al.*, 2020b), R2GenCMN (Chen *et al.*, 2021)) and models introducing external information (PPKED (Liu *et al.*, 2021a), MSAT (Wang *et al.*, 2022c)). On MIMIC-CXR, METransformer is the best performer across all metrics. Especially, our CIDEr score is up to 0.362, which is to date the best performance and makes a significant improvement over the second best score of

#	Models	IU-Xray					MIMIC-CXR				
		BLEU_4	ROUGE	METEOR	CIDEr	AVG. Δ	BLEU_4	ROUGE	METEOR	CIDEr	AVG. Δ
1	BASELINE	0.161	0.357	0.183	0.337	-	0.109	0.277	0.143	0.275	-
2	+BE	0.163	0.360	0.185	0.346	+1.5%	0.111	0.279	0.144	0.287	+1.9%
3	+BE+ETs	0.168	0.372	0.188	0.402	+7.6%	0.117	0.284	0.147	0.334	+8.5%
4	+BE+ETs+OrL	0.170	0.377	0.190	0.426	+10.4%	0.121	0.287	0.150	0.352	+11.9%
5	+BE+ETs+OrL+EV	0.172	0.380	0.192	0.435	+11.8%	0.124	0.291	0.152	0.362	+14.2%

Table 7.3. Ablation studies. The “BASELINE” uses ViT encoder and bilinear-attention decoder with the traditional cross entropy-loss. Here “BE”, “ETs”, “OrL”, “EV” stand for “Bilinear-attention Encoder”, “Expert Tokens”, “Orthogonal Loss”, and “Expert Voting”. The number of expert tokens used in the experiments for the IU-xray and mimic-cxr datasets were 5 and 7, respectively. The average improvement over all NLG metrics compared to baseline is also presented in the “AVG. Δ ” column.

0.299 of MSAT (Wang *et al.*, 2022c). These improvements demonstrate the advantage of our framework which is conceptually “multiple expert co-diagnosis”. Ours METransformer also surpasses these methods on IU-Xray over most metrics, while it is again worth mentioning that the cited performances on IU-Xray may not be strictly comparable as obscure training-test partitions were used by different methods.

7.3.4 Ablation Study

Contribution of each component. We conduct an ablation study to single out the contributions of each model component, as presented in Table 7.3. We build a baseline by using ViT transformer encoder and bilinear-attention decoder to verify the performance improvements brought by multiple expert tokens, orthogonal loss, and our metric-based expert voting strategy. In Table 7.3, there are four components: “BE”, “ETs”, “OrL”, and “EV”, representing Bilinear-attention Encoder, Expert Tokens, Orthogonal Loss, and Expert Voting, respectively. The symbol “+” indicates the inclusion of the following component based on the “BASELINE” model. It can be observed that each component of METransformer has a positive effect on performance. By comparing #3 and #1 in Table. 7.3, it can be found that extending the expert tokens and enhancing them with bilinear-attention encoder can increase the overall performance by 7.6% on

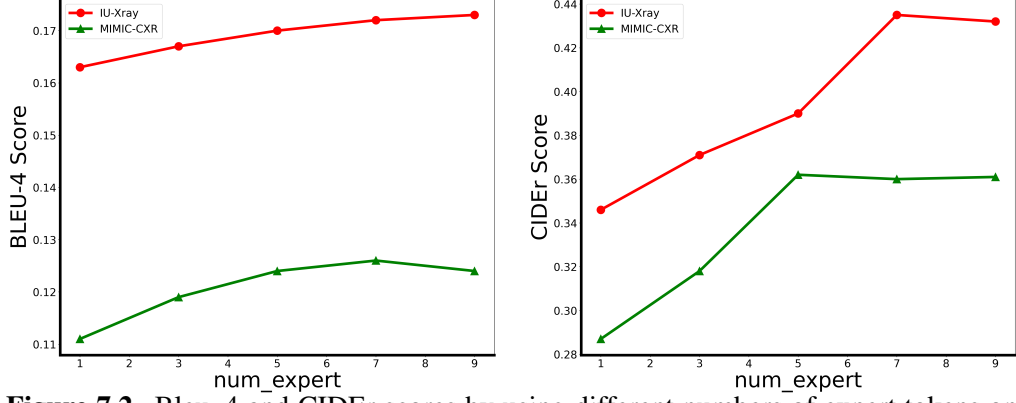


Figure 7.2. Bleu_4 and CIDEr scores by using different numbers of expert tokens on IU-Xray and MIMIC-CXR dataset.

IU-Xray and 8.5% on MIMIC-CXR. Training with the orthogonal loss on expert tokens can further enhance model performance (by comparing #4 and #3). For experiments #3 and #4 where multiple expert tokens are used, the final prediction result is obtained by averaging the probability of words predicted by the multiple experts. Comparing #5 and #4 shows that our voting strategy is more effective than the averaging method above.

Impact of expert tokens. To show the impacts of the expert tokens, we train METransformer with different numbers of expert tokens, i.e., $num_expert \in \{1, 3, 5, 7, 9\}$ and the results on IU-Xray and MIMIC-CXR are shown in Figure. 7.2. We observe the following. First, increasing the expert tokens can significantly improve the overall performance of the model. This validates the effectiveness of our motivation that by focusing on different image regions with multiple experts, the model can learn more diversified information and thus produce more accurate and diverse diagnostic reports. Second, when the number of expert tokens exceeds a threshold, increasing num_expert is not able to continue promising a better outcome, for example, when num_expert is increased from 7 to 9, the CIDEr score on the IU-Xray dataset decreases. A possible explanation is that our model forces each expert token to focus on a different image region (via orthogonal loss), which may cause some experts to attend to irrelevant image regions when there are too many experts, thus may negatively affect the generation process.

Models	Params	Training Time	BLEU_4	ROUGE	METEOR	CIDEr
METransformer(num_expert=1)	152M	0.48h	0.163	0.362	0.183	0.346
Stochastic Model Averaging	152M x 7=1064M	0.48h x 7=3.36h	0.168	0.373	0.187	0.389
Random Initialization	152M x 7=1064M	0.48h x 7=3.36h	0.170	0.376	0.189	0.408
Multiple Decoder	483.3M	1.85h	0.166	0.368	0.186	0.378
METransformer(num_expert=7)	152.007M	0.55h	0.172	0.380	0.192	0.435

Table 7.4. Comparison with the ensemble models on the IU-Xray dataset. The efficiency results refer to the total time required to complete the training using one RTX 3090.

Comparison with ensemble models. By using multiple experts, our model is conceptually analogous to an ensemble model. We thus compare with three ensemble models in Table. 7.4. “Random Initialization” trains randomly initialized METransformer(num_expert=1) model for 7 times, corresponding to random-seed based ensemble. “Multiple Decoder” trains the model with the encoder of METransformer(num_expert=1) and 7 randomly initialized METransformer’s decoders, corresponding to late fusion. We also compare with a stochastic model averaging ensemble method. All methods ensemble the results using our proposed Expert Voting strategy. As observed, our METransformer(num_expert=7) performs significantly better with much fewer trainable parameters. We attribute this to the sophisticated interaction between expert tokens through our compact design. Compared with using a single expert token, our method using 7 expert tokens only incurs 0.007M (0.05%) extra parameters, demonstrating its scaling ability.

7.3.5 Qualitative analysis.

Visualisation of expert tokens. To gain insight, we visualize the image regions mostly attended by each learned expert token in Figure. 7.3 via exploring the attention $\hat{\alpha}_s$ between the learned expert token embeddings $\hat{\mathbf{z}}_L^e$ and the visual token embeddings $\hat{\mathbf{z}}_L^v$: $\hat{\alpha}_s = \text{Softmax}(\hat{\mathbf{z}}_L^e(\hat{\mathbf{z}}_L^v)^T)$

Using $\text{Softmax}(\cdot)$, only the mostly attended image regions are shown. As observed, each expert token attends to a distinct and critical image region. For

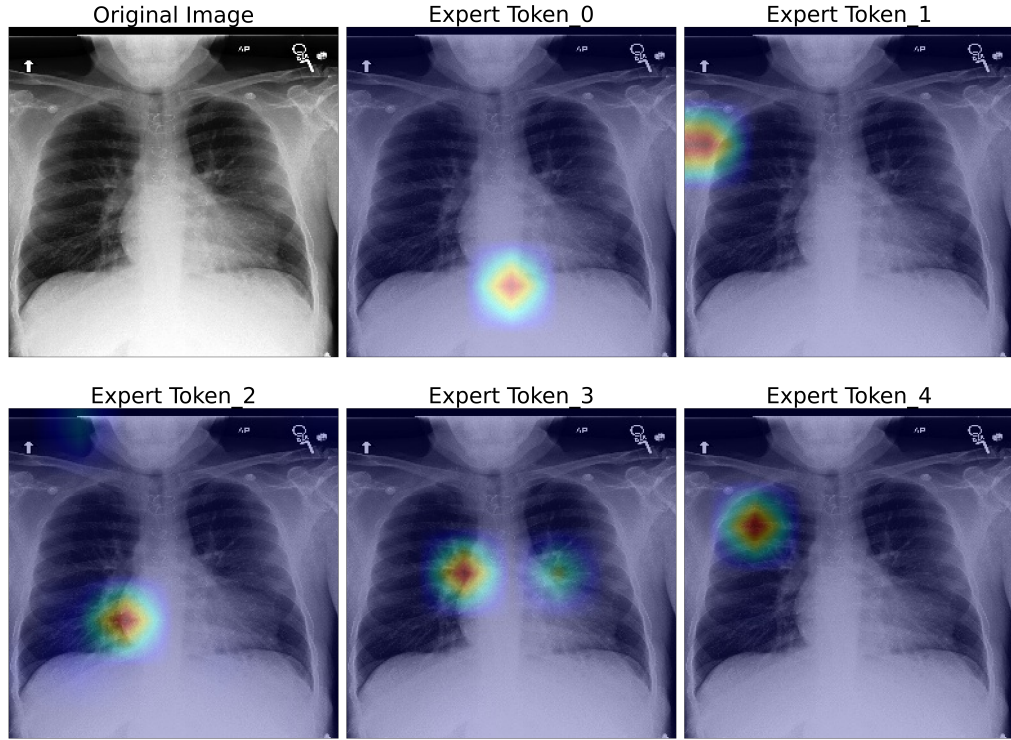


Figure 7.3. Attention visualization of expert tokens on the image.

example, the image region attended by expert Token_2 is known as the angle of the rib diaphragm which can provide valuable clinic information, e.g., when the angle of the rib diaphragm is not sharp enough or is even obtuse, it suggests a small amount of pleural effusion.

Qualitative results. We show the generated results of METransformer compared with the Baseline method (“#1” in Table. 7.3) in Figure. 7.4. On the left, we visualize the image-text attention mapping from the last layer of the expert transformer decoder with three key generated medical terms. As observed, METransformer better aligns the locations with the related text. On the right, we present the corresponding generated reports and the ground truth. For better illustration, we differently color sentences containing different medical information. It is observed that METransformer is able to generate descriptions better aligned with that written by radiologists. For example, METransformer can diagnose anomalies in the heart part, while the Baseline model fails. This is

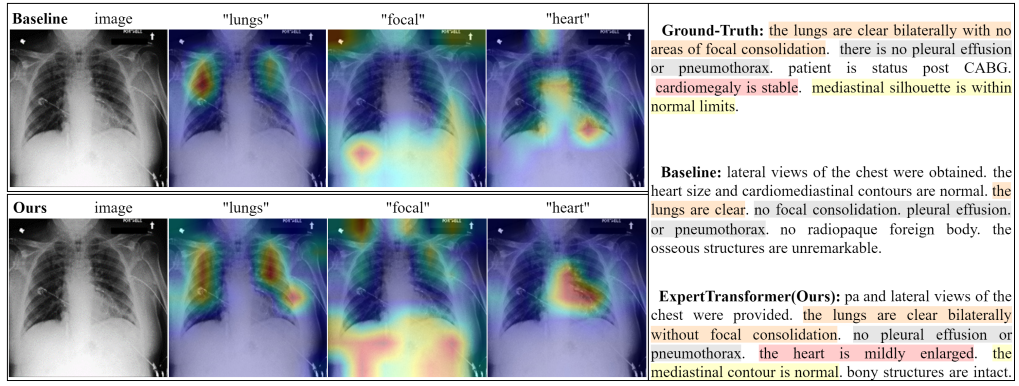


Figure 7.4. An example of the generated reports and their attention-mapping visualization of three key medical terms from BASELINE and ours METransformer. For better illustration, different colors in the generated reports highlight different medical information.

consistent with the fact that our model can better attend to the heart part of the image (see the attended area of “heart” on the left side of the figure).

7.4 Conclusions

We present an effective approach for radiology report generation from a new perspective orthogonal to existing research efforts in this field. Our approach follows the concept of multi-specialist consultation to improve the quality of generated reports by introducing multiple learnable expert tokens into a transformer-base framework. Despite its promising performance and properties demonstrated in the experiment, our METransformer is still **limited** in being a basic framework that could be further enhanced by integrating medical domain knowledge, as seen in “single-expert” based methods, which will be explored in our future work.

R2GenGPT: Radiology Report Generation with Frozen LLMs

Large Language Models (LLMs) have consistently showcased remarkable generalization capabilities when applied to various language tasks. Nonetheless, harnessing the full potential of LLMs for Radiology Report Generation (R2Gen) still presents a challenge, stemming from the inherent disparity in modality between LLMs and the R2Gen task. To bridge this gap effectively, we propose R2GenGPT, which is a novel solution that aligns visual features with the word embedding space of LLMs using an efficient visual alignment module. This innovative approach empowers the previously static LLM to seamlessly integrate and process image information, marking a step forward in optimizing R2Gen performance. R2GenGPT offers the following benefits. First, it attains state-of-the-art (SOTA) performance by training only the lightweight visual alignment module while freezing all the parameters of LLM. Second, it exhibits high training efficiency, as it requires the training of an exceptionally minimal number of parameters while achieving rapid convergence. By employing delta tuning, our model only trains 5M parameters (which constitute just 0.07% of the total parameter count) to achieve performance close to the SOTA levels.

8.1 Introduction

Given its critical clinical relevance, the field of radiological report generation (R2Gen) has been garnering increasing attention. As we have detailed in the review [2.2](#), recent works in the R2Gen task mainly aim to tackle two major

challenges. The first challenge lies in **long text generation**. Unlike the image captioning task which generates a single sentence description, medical report generation requires detailed and coherent paragraph-long descriptions. This requires the model to have a robust capacity for learning long-range dependencies. To address this, many solutions have been proposed (Jing *et al.*, 2018; Yin *et al.*, 2019; Wang *et al.*, 2021; Chen *et al.*, 2020b; Chen *et al.*, 2021). For instance, some research works (Jing *et al.*, 2018; Yin *et al.*, 2019; Wang *et al.*, 2021) have employed hierarchically structured LSTM which first produces topic vectors using a sentence LSTM and then creates a description for each generated topic with a word LSTM. Another type of work R2Gen (Chen *et al.*, 2020b) introduced a memory-driven Transformer that can record key information of the generation process, enhancing the model’s ability to produce long texts. The second challenge lies in the **bias in visual and textual data**. Due to an over-representation of normal samples in the training data, the model’s learning process was biased towards these samples, limiting its ability to effectively detect abnormalities and anomalies within the dataset. Some works (Wang *et al.*, 2021; You *et al.*, 2021; Wang *et al.*, 2022b) have addressed this issue by aligning image and text/report features, such as work Self-boost (Wang *et al.*, 2021) incorporating an image-text matching branch to enhance the model’s capability to capture the anomalous features in the image. Other research works mitigate the effects of data bias by incorporating external knowledge, such as medical tags (Jing *et al.*, 2018; Wang *et al.*, 2022c), and knowledge graphs (Li *et al.*, 2019a; Zhang *et al.*, 2020; Liu *et al.*, 2021a; Yang *et al.*, 2021b). For instance, PPKED (Liu *et al.*, 2021a) utilizes a knowledge graph and introduces the Posterior-and-Prior Knowledge Exploring-and-Distilling framework. Despite many efforts and solutions putting forth, the aforementioned two challenges remain significant issues in this field.

Recently, large language models (LLMs) (e.g., (Chowdhery *et al.*, 2022; Touvron *et al.*, 2023)) have demonstrated excellent capabilities to perform tasks

with zero in-domain data, conduct logical reasoning, and apply commonsense knowledge in NLP tasks (Kojima *et al.*, 2022; Wei *et al.*, 2022). This leads us to ponder whether we can apply large language models to medical report generation tasks, as pre-trained large language models seem to inherently possess the ability to address the two challenges mentioned above. As for long text generation, LLMs are equipped with an inherent understanding of grammar, syntax, and semantic coherence, making them well-suited for tasks requiring extended text generation, such as medical reporting. Furthermore, their proficiency in context modeling allows them to maintain consistency and relevance throughout a lengthy report. As for the bias stemming from an over-representation of normal samples in medical datasets, LLMs can serve as potential correctives due to their extensive knowledge base. Having been exposed to vast amounts of data, LLMs demonstrate robustness and are less susceptible to the effects of imbalanced datasets. They are even capable of handling numerous zero-shot tasks. Moreover, current methods mitigating bias entail the incorporation of external knowledge, whereas pre-trained LLMs inherently possess a wealth of informative knowledge.

However, applying LLMs to R2Gen tasks poses challenges due to the fundamental disparity between visual and textual modalities. The crucial step in applying LLMs to R2Gen is to bridge the gap between visual information and textual generation. In this chapter, we present R2GenGPT and explore three methods for aligning visual features with large language models. We first process chest x-ray images using a Visual Encoder to obtain visual embeddings. These embeddings are then mapped to the LLM’s feature space via a Visual Mapper, ensuring uniform dimensions. To identify the most efficient method of aligning visual features with the LLM, we’ve crafted three alignment modules: 1) shallow alignment, where only the Visual Mapper is trained and other parameters remain fixed; 2) deep alignment, where both the visual encoder and the Visual Mapper are trained simultaneously; and 3) Delta alignment, where

the Visual Mapper and a limited set of incremental parameters from the visual encoder are trained, ensuring both effectiveness and efficiency.

Our main contributions are summarized as follows.

- We propose a novel LLMs-based Radiology report generation (R2Gen) framework, dubbed R2GenGPT. This marks the first instance of harnessing pre-trained large language models (LLMs) for the R2Gen task with comprehensive comparisons conducted on two frequently employed benchmark datasets.
- We explored three methods with varying levels of trainable parameters to connect image modalities to large language models, namely: shallow alignment, delta alignment, and deep alignment, enabling the LLM to effectively process visual information.
- Our approach exhibits promising and robust performance on two widely recognized benchmark datasets—IU-Xray and MIMIC-CXR. In comparison to multiple state-of-the-art methods, our framework consistently demonstrates its efficacy, affirming its potential in the field of R2Gen.

8.2 Methodology

Overview As illustrated in Figure 8.1, R2GenGPT comprises a Visual Encoder, a Visual Mapper, and an LLM (Large Language Model) component. The visual encoder is employed to extract information from chest x-ray images, while the visual mapper serves to project low-dimensional image features into the high-dimensional feature space of the LLM. Utilizing the visual features derived from the chest x-ray images, the LLM generates corresponding diagnostic reports.

Feature Alignment For an input chest xray image X_v , we consider the pre-trained Swin Transformer (Liu *et al.*, 2021c) as visual encoder, which provides

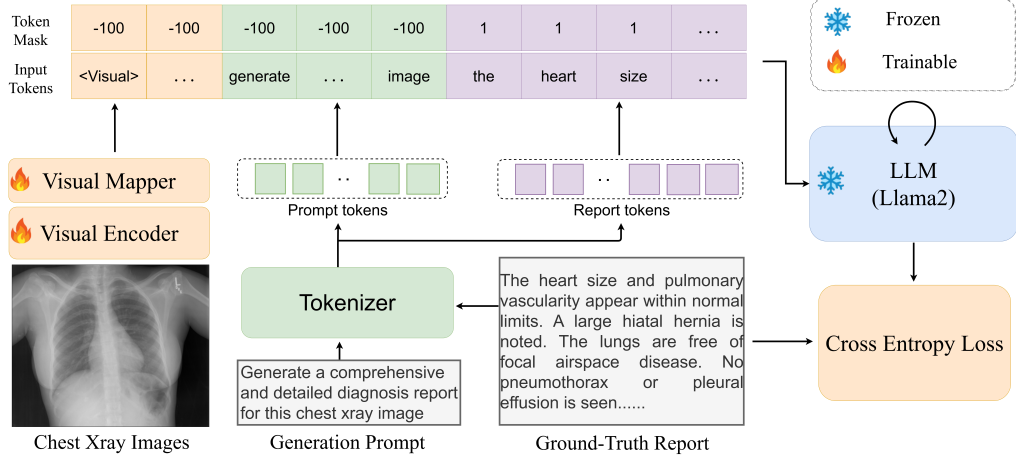


Figure 8.1. An overview of our proposed R2GenGPT. The input tokens for the Large Language Model (LLM) are sequentially concatenated, consisting of visual tokens, prompt tokens, and report tokens. A token mask of -100 indicates that those particular tokens are excluded from auto-regressive training, while a mask of 1 signifies inclusion in auto-regressive training.

the visual feature $\mathbf{Z}_v = g(\mathbf{X}_v; \theta_v)$, where θ_v is the parameters of the Swin Transformer. The grid features of the last transformer layer is utilized in our experiments. We consider a simple linear layer as the Visual Mapper to connect image features into the LLM’s word embedding space. Specifically, we apply a trainable projection matrix $\mathbf{W}_m$ to convert $\mathbf{Z}_v$ into language embedding tokens $\mathbf{H}_v$, which have the same dimensionality of the word embedding space in the large language model.

$$\mathbf{H}_v = \mathbf{W}_m \mathbf{Z}_v, \quad \text{with } \mathbf{Z}_v = g(\mathbf{X}_v) \quad (8.1)$$

Thus we have a sequence of visual tokens $\mathbf{H}_v$. Following the extraction of visual tokens $\mathbf{H}_v$, we propose the following three distinct training strategies (as illustrated in Figure 8.2) to identify the most efficient aligning method by varying the level of trainable parameters.

- a) Shallow Alignment: In this mode, we fix the parameters of the pre-trained Swin Transformer and train only the linear Visual Mapper, represented by $\mathbf{W}_m$.
- b) Deep Alignment: For this approach, both the Swin Transformer and the Visual Mapper are jointly fine-tuned. Specifically, parameters from both the Visual

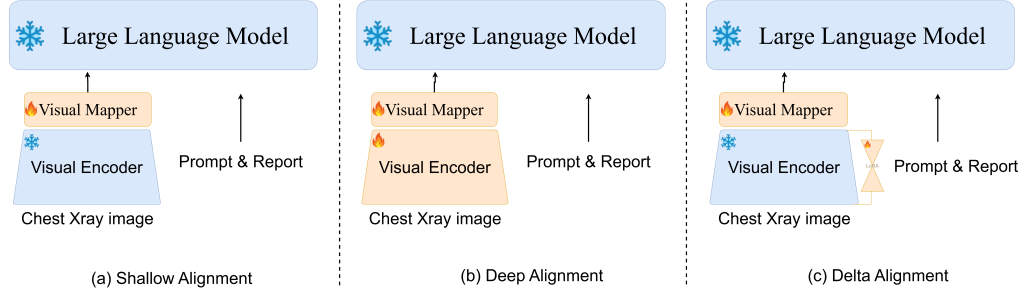


Figure 8.2. Three proposed alignment methods. (a) Shallow Alignment: Training only the Linear Layer. (b) Deep Alignment: Training both the Linear Layer and all parameters of the Visual Encoder. (c) Delta Alignment: Training the Linear Layer and a small subset of incremental parameters of the Visual Encoder.

Encoder (Swin Transformer) and the Visual Mapper, denoted as θ_v and $\mathbf{W}_m$ respectively, are updated.

c) Delta Alignment: As the Swin Transformer utilized in this chapter was originally trained on natural images, the shallow alignment approach hinders the model’s ability to capture high-quality radiographic image features. On the other hand, adopting deep alignment substantially impacts the model’s training efficiency. Therefore, we propose delta alignment, parameter-efficiently fine-tuning the Swin Transformer model using LoRA (Hu *et al.*, 2022). Specifically, for a pre-trained weight matrix $\mathbf{W}_0$ within θ_v , LoRA constrains its update with two smaller matrices using a low-rank decomposition $\mathbf{W}_0 + \Delta\mathbf{W}_0 = \mathbf{W}_0 + \mathbf{BA}$, where $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$, $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, and the rank $r \ll \min(d, k)$. It is noted that in our implementation, we only adjust the query and value projections within the swin transformer to prioritize a simple yet efficient model. The trained parameters are denoted as $\Delta\theta_v$, and both $\Delta\theta_v$ and $\mathbf{W}_m$ are trained in this mode.

Large Language Models We adopt Llama2-7B model for the large language model component. The Llama2-7B stands out for its remarkable capabilities and robustness. Designed with a massive 7-billion-parameter architecture, it encapsulates a rich knowledge base derived from extensive pre-training on diverse datasets. One of its key strengths lies in its extraordinary ability to understand

and generate complex language structures, making it particularly well-suited for intricate tasks such as radiology report generation.

Given an chest xray image $\mathbf{X}_v$ and its corresponding report $\mathbf{X}_r$, the detailed prompt inputted into Llama2 is as follows.

$$\text{Human: } \langle \text{Img} \rangle \mathbf{X}_v \langle / \text{Img} \rangle, \mathbf{X}_p. \backslash \text{n Assistant: } \mathbf{X}_r \langle / s \rangle.$$

Here $\mathbf{X}_p$ is our designed instruction prompt specific to the R2Gen task. In our current implementation, $\mathbf{X}_p$ = “Generate a comprehensive and detailed diagnosis report for this chest xray image.”. For this prompt, before inputting it into LLAMA2 for computation, $\mathbf{X}_v$ will be replaced by visual tokens $\mathbf{H}_v$ processed using Equ. 8.1 while all other text is tokenized into word tokens using LLAMA’s tokenizer.

Loss Function We perform instruction-tuning of the LLM only on the report tokens , using its original auto-regressive training objective. Specifically, for a report of length L , conditioned on visual information $\mathbf{X}_v$ and instruction prompt $\mathbf{X}_p$, our loss function, captured as the negative log likelihood, is formulated as:

$$\mathcal{L}(\theta; \mathbf{X}_r, \mathbf{X}_v, \mathbf{X}_p) = - \sum_{i=1}^L \log p_{\theta}(x_i | \mathbf{X}_v, \mathbf{X}_p, \mathbf{X}_{r, < i}), \quad (8.2)$$

where θ is the trainable parameters, $\mathbf{X}_{r, < i}$ is the report tokens before the current prediction token x_i .

8.3 Experiments

8.3.1 Data Collection

We evaluated performance using two datasets: a widely-used benchmark IU-Xray (Demner-Fushman *et al.*, 2016) and the currently largest dataset MIMIC-CXR (Johnson *et al.*, 2019) for medical report generation.

IU-Xray: Indiana University Chest X-ray Collection (IU-Xray) (Demner-Fushman *et al.*, 2016) is the most widely used publicly accessible dataset in medical report generation tasks. It contains 3955 fully de-identified radiology reports, each of which is associated with frontal and/or lateral chest X-ray images, and 7470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings, Indication, etc. In this work, we adopt the same data set partitioning as (Chen *et al.*, 2020b) for a fair comparison, with a train/test/val set by 7:1:2 of the entire dataset. All evaluations are done on the test set.

MIMIC-CXR: This largest publicly available dataset encompasses both chest radiographs and unstructured textual reports. This comprehensive dataset comprises a total of 377110 chest X-ray images and 227835 corresponding reports sourced from 64588 patients who underwent examination at the Beth Israel Deaconess Medical Center between 2011 and 2016. To ensure equitable comparisons, we adhered to MIMIC-CXR’s official partitioning as outlined in (Chen *et al.*, 2020b), resulting in 270790 samples designated for training, while allocating 2130 and 3858 samples for validation and testing, respectively.

8.3.2 Experimental Settings

Evaluation Metrics Adhering to the established evaluation protocol¹, we employ the prevalent metrics for assessment, namely BLEU scores (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and

¹<https://github.com/tylin/coco-caption>

Table 8.1. Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from their respective papers. For the methods without †, their results are obtained by re-running the publicly released codebase (Li *et al.*, 2021) on these two datasets using the same training-test partition as our method.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
IU-Xray	Show-Tell	0.243	0.130	0.108	0.078	0.307	0.157	0.197
	Att2in	0.248	0.134	0.116	0.091	0.309	0.162	0.215
	AdaAtt	0.284	0.207	0.150	0.126	0.311	0.165	0.268
	Transformer	0.372	0.251	0.147	0.136	0.317	0.168	0.310
	M2transformer	0.402	0.284	0.168	0.143	0.328	0.170	0.332
	R2Gen [†]	0.470	0.304	0.219	0.165	0.371	0.187	-
	R2GenCMN [†]	0.475	0.309	0.222	0.170	0.375	0.191	-
	MSAT [†]	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	METransformer [†]	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	R2GenGPT (Shallow)	0.466	0.301	0.211	0.156	0.370	0.202	0.405
	R2GenGPT (Delta)	0.470	0.299	0.213	0.162	0.369	0.211	0.419
	R2GenGPT (Deep)	0.488	0.316	0.228	0.173	0.377	0.211	0.438
MIMIC-CXR	Show-Tell	0.308	0.190	0.125	0.088	0.256	0.122	0.096
	Att2in	0.314	0.198	0.133	0.095	0.264	0.122	0.106
	AdaAtt	0.314	0.198	0.132	0.094	0.267	0.128	0.131
	Transformer	0.316	0.199	0.140	0.092	0.267	0.129	0.134
	M2Transformer	0.332	0.210	0.142	0.101	0.264	0.134	0.142
	R2Gen [†]	0.353	0.218	0.145	0.103	0.277	0.142	-
	R2GenCMN [†]	0.353	0.218	0.148	0.106	0.278	0.142	-
	PPKED [†]	0.36	0.224	0.149	0.106	0.284	0.149	0.237
	GSK [†]	0.363	0.228	0.156	0.115	0.284	-	0.203
	MSAT [†]	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	METransformer [†]	0.386	0.250	0.169	0.124	0.291	0.152	0.362
	R2GenGPT (Shallow)	0.365	0.237	0.163	0.117	0.277	0.136	0.145
	R2GenGPT (Delta)	0.380	0.244	0.167	0.119	0.281	0.145	0.195
	R2GenGPT (Deep)	0.411	0.267	0.186	0.134	0.297	0.160	0.269

CIDEr (Vedantam *et al.*, 2015), to gauge the quality of the generated textual reports. To measure the accuracy of descriptions for clinical abnormalities, we follow (Chen *et al.*, 2020b; Chen *et al.*, 2021; Liu *et al.*, 2021b) and further report clinical efficacy metrics. Specifically, we employ CheXpert (Irvin *et al.*, 2019) for annotating the generated reports, which are subsequently compared against ground truth annotations across 14 distinct categories related to thoracic

diseases and support devices. We use precision, recall, and F1 to evaluate model performance for clinical efficacy metrics.

Implementation Details In this work, we leveraged the LLAMA2-7B model ² as the large language model and the base version of the Swin Transformer ³ as the Visual Encoder. Within the parameters of LoRA, we configured the Lora attention dimension to 16, and the alpha parameter for Lora scaling was also set at 16. The training process was conducted on four NVIDIA A100 40GB GPUs using mixed precision for 3 epochs for MIMIC-CXR and 15 epochs for IU-Xray dataset, with a mini-batch size of 6 and a learning rate of 1e-4. During the testing phase, we employed a beam search strategy with a beam size set to 3.

8.3.3 Results and Discussion

Comparison with SOTA Table 8.1 showcases a performance comparison between the state-of-the-art methods and our R2GenGPT model variants on the IU-Xray and MIMIC-CXR dataset. In terms of standard image captioning methods, the table considers Show-Tell (Vinyals *et al.*, 2015), Att2in (Xu *et al.*, 2015), AdaAtt (Lu *et al.*, 2017), Transformer (Vaswani *et al.*, 2017), and M2Transformer (Cornia *et al.*, 2020). Furthermore, medical report generation methods such as R2Gen (Chen *et al.*, 2020b), R2GenCMN (Chen *et al.*, 2021), MSAT (Wang *et al.*, 2022c), METransformer (Wang *et al.*, 2023), and other methods marked with † in Table 8.1 are considered.

From Table 8.1, it is evident that our R2GenGPT model variants, especially R2GenGPT (Deep), outperform the compared methods across nearly all evaluation metrics. In the MIMIC-CXR dataset, apart from CIDEr, we significantly outperform the latest METransformer (Wang *et al.*, 2023) method across all metrics. For instance, our BLEU_4 score is improved from 0.124 to 0.134, marking

²<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

³<https://huggingface.co/microsoft/swin-base-patch4-window7-224>

Table 8.2. Evaluation of Model Efficiency and Clinical Efficacy on MIMIC-CXR dataset.

Models	Trainable Components			Scale and Efficiency		Clinical Efficacy		
	Mapper	Encoder	LoRA	Trainable Parameter	Time	Precision	Recall	F1
Shallow	✓			4.2M	1.75h/epo	0.351	0.291	0.302
Delta	✓	✓		5.0M	1.83h/epo	0.366	0.305	0.316
Deep	✓		✓	90.9M	2.75h/epo	0.392	0.337	0.342
METransformer	-	-	-	152M	3.62h/epo	0.364	0.309	0.311

an 8.1% increase. However, we achieved a CIDEr score of 0.269, which is lower than METransformer’s 0.362. This discrepancy is because METransformer employs an expert voting strategy similar to an ensemble approach to enhance the CIDEr metric. In comparison to methods without this enhancement, such as R2Gen (Chen *et al.*, 2020b) and PPKED (Liu *et al.*, 2021a), we also hold a distinct advantage in terms of the CIDEr metric. It is also noteworthy that our R2GenGPT (Shallow), with only 4.2M trainable parameters, has been able to achieve a performance in par with the well-known R2Gen model (Chen *et al.*, 2020b), if not even better.

Model Efficiency and Clinical Efficacy Analysis In Table 8.2, we have presented both model efficiency and clinical efficacy metrics. It’s evident that our model exhibits higher training efficiency compared to METransformer. For instance, R2GenGPT (Deep) requires training with only 90.9 million parameters, which is significantly less than METransformer’s 152 million. Furthermore, R2GenGPT (Delta) achieves comparable performance with just 5 million parameters. To assess the model’s training efficiency, we conducted evaluations on four A100 40G GPUs and recorded the time required for one training epoch. Notably, R2GenGPT Shallow, Delta, and Deep each completed this epoch in just 1.75, 1.83, and 2.75 hours, respectively, compared to METransformer’s 3.62 hours, highlighting our model’s superior training efficiency. In terms of clinical efficacy metrics, it can be observed that our R2GenGPT(deep)

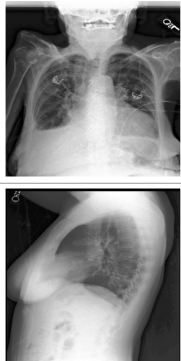
	Ground Truth	Shallow Align	Delta Align	Deep Align
	impression : mild pulmonary edema with small bilateral pleural effusions and mild bibasilar atelectasis . findings : the patient is status post median sternotomy and aortic valve replacement . the heart size is moderately enlarged but unchanged . there is mild pulmonary edema new from the prior study . small bilateral pleural effusions are present .	impression : no acute cardiopulmonary process . findings : the patient is status post median sternotomy and coronary artery bypass grafting . the heart size is mildly enlarged . there is no pleural effusion or pneumothorax . there is no focal consolidation concerning for pneumonia . the lungs are otherwise clear . the mediastinal contours are unremarkable .	impression : mild pulmonary edema with small bilateral pleural effusions . findings : the patient is status post median sternotomy and coronary artery bypass graft surgery . there is mild pulmonary edema with small bilateral pleural effusions . there is no pneumothorax . the cardiomeastinal silhouette is unchanged . sternal wires are intact . a right-sided port-a-cath terminates in the mid svc .	impression : mild pulmonary edema with small bilateral pleural effusions . findings : the patient is status post median sternotomy cabg and aortic valve replacement . the heart size remains moderately enlarged . there is mild pulmonary edema with small bilateral pleural effusions . there is no pneumothorax . bony structures are intact . degenerative changes are noted in the thoracic spine .
	Ground Truth	Shallow Align	Delta Align	Deep Align
	impression : no acute cardiopulmonary process . findings : cardiac mediastinal and hilar contours are normal . no focal consolidation pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .	impression : no acute cardiopulmonary process . findings : pa and lateral views of the chest provided . there is no focal consolidation pleural effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .	impression : no acute cardiopulmonary abnormality . findings : heart size is normal . the mediastinal and hilar contours are normal . the pulmonary vasculature is normal . lungs are clear . no pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities . surgical clips project over the right upper quadrant of the abdomen .	impression : no acute cardiopulmonary abnormality . findings : heart size is normal . the mediastinal and hilar contours are normal . the pulmonary vasculature is normal . lungs are clear . no pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities . mild degenerative changes are noted in the thoracic spine .

Figure 8.3. Examples of the generated report on MIMIC-CXR dataset. We compared the results generated by the three alignment methods proposed in R2GenGPT. For better illustration, the key medical information in the reports are highlighted using different colors.

and R2GenGPT(delta) achieved F1 scores of 0.342 and 0.316, respectively, surpassing the current SOTA method, METransformer, with a score of 0.311. This demonstrates the ability of R2GenGPT to generate crucial clinical information.

Qualitative results In Figure 8.3, we compare the reports generated by the three alignment methods of R2GenGPT. To provide a better visualization, we highlight key medical information in both the ground truth and generated reports using different colors. From the figure, it can be observed that reports generated using the Shallow Alignment method are notably inferior to those generated using the Delta Alignment and Deep Alignment methods. For instance, in the first example (top), the Shallow Alignment method erroneously identifies a sample with mild pulmonary edema as a normal sample, whereas Delta Alignment and Deep Alignment methods can accurately identify it.

8.4 Conclusions

In this chapter, we present R2GenGPT, an innovative framework at the forefront of Radiology Report Generation (R2Gen) that capitalizes on the capabilities of Large Language Models (LLMs). Through a comprehensive exploration

of three alignment methods—shallow, delta, and deep— this research highlights the game-changing potential of LLMs in elevating the R2Gen landscape. R2GenGPT not only attains competitive SOTA performance but also achieves a remarkable reduction in computational complexity. This dual achievement positions R2GenGPT as a promising solution to automate and improve radiology reporting.

Conclusions

In this thesis, we have orchestrated a comprehensive journey that systematically advanced the field of automated radiology report generation through a series of progressive methodological innovations. Our work can be conceptualized as a multi-staged endeavor that evolved in complexity and efficiency at each phase.

Overall, in this thesis, to address the challenges faced in Radiology Report Generation, we have proposed multiple strategies: i) introducing auxiliary tasks such as image-text matching and image classification to enhance the model’s ability to capture fine-grained critical visual feature differences, thereby improving the quality of report generation (Chapters 3, 4, and 6), ii) incorporating external knowledge information such as medical concepts and medical knowledge graphs to enrich the semantic content of the generated reports (Chapter 5), iii) proposing a multi-expert joint diagnostic framework to enhance the quality and accuracy of the generated reports (Chapter 7) and utilizing the powerful world knowledge and language comprehension abilities of large language models (LLMs) (Chapter 8). These approaches aim to enhance the quality of model-generated reports from various perspectives, but essentially, they all address the challenges faced in radiology report generation tasks, as we introduced in the introduction.

Specifically, in the initial stages, outlined in Chapter 3, we developed a self-boosting framework that had a twofold impact. First, it seamlessly combined the auxiliary task of image-text matching with medical report generation. Second,

the method leveraged a new loss term for evaluating feature similarity between generated and ground-truth reports, contributing to fine-grained diagnostic capabilities. The result was a tightly-coupled, mutually reinforcing approach that significantly improved the accuracy of diagnostic reports. Building upon this foundation, Chapter 4 introduced a sample-graph consistency loss to enforce uniformity in visual and textural feature embeddings. This extension of the self-boosting framework brought further granularity to diagnostic captioning, acting as a catalyst in the refinement of report quality. By Chapter 5, the narrative shifted to the application of advanced attention mechanisms. Here, we deployed a bilinear-pooling-assisted sparse attention block within a transformer network. This innovation heightened the sensitivity to fine-grained visual variations in radiographic images, while an additional medical concepts generation network enriched the semantic content of the generated reports. Chapter 6 marked a radical departure from convention by pioneering the use of a pure transformer model, free from CNN constraints. By applying a multi-criteria supervision scheme—incorporating an image-text matching loss for semantic coherence and a term-weighting scheme with multi-label classification—we were able to capture critical diagnostic information without adding unnecessary complexity to the transformer structure. In Chapter 7, we introduced a groundbreaking framework, the METransformer, designed to simulate the collaborative environment of multi-expert joint diagnoses. By incorporating learnable expert tokens and utilizing both linear and non-linear attention mechanisms, this framework synthesized diverse knowledge representations. Significantly, the METransformer not only exceeded the performance of conventional ensemble methods but also achieved this while being more computationally efficient. The crescendo of our work came with the introduction of R2GenGPT in Chapter 8. This framework leveraged pre-trained large language models, fortified with three novel methods for visual information integration: shallow alignment, delta alignment, and deep alignment. This not only set new benchmarks in performance but also opened

doors for future work in integrating large language models into medical informatics.

In conclusion, the scope and impact of the work encapsulated in this thesis serve as both a benchmark and a foundation for future research. The methodologies developed herein not only advance the state-of-the-art but also offer promising avenues for continued innovation in automated radiology report generation.

9.1 Future Outlook

Despite the significant progress achieved, numerous promising avenues for research remain open. Each area presents unique challenges and opportunities for further innovation.

- **Long-Tail Distribution of Terms:** Currently, the issue of the long-tail distribution of terms in medical reports is underexplored. Addressing this could involve the development of specialized loss terms that consider class imbalance. Techniques like oversampling rare classes, or employing weighted loss functions could make models more sensitive to less frequent but potentially important terms.
- **Computational Efficiency:** While the efficacy of transformers in medical report generation has been proven, their computational requirements remain a bottleneck. One direction for future work could be exploring lightweight transformer architectures or employing model-pruning techniques to reduce computational load without sacrificing performance.
- **Inclusion of Medical Domain Knowledge:** Integrating structured medical domain knowledge via frameworks like knowledge graphs presents an exciting avenue for improvement. Such integration could be accomplished through multi-modal learning algorithms that can process both

textual and structured data, thereby offering more nuanced and contextually rich reports.

- **Optimization of Large Language Models (LLMs):** Our initial forays into employing LLMs have shown promise, but there's much room for improvement. Potential approaches could include fine-tuning these models with a domain-specific corpus or developing custom training regimes to better adapt them to the specialized requirements of radiology reporting. Furthermore, the methodology of this thesis primarily focuses on the generation of medical reports in English. Therefore, another avenue for improvement lies in expanding the generation of medical reports to various language formats using multilingual large language models. This expansion would significantly enhance the practical applicability of the model.
- **Localized Evaluation:** In future work, we aim to assess the effectiveness of our methods using data from local patients. Collaborating with regional healthcare providers, we will conduct a more targeted evaluation, ensuring the applicability of our findings to the specific characteristics and healthcare practices of the community. This localized approach will enhance the overall effectiveness and relevance of our methods in real-world healthcare settings.
- **Multimodal Data Integration:** Future work could also focus on integrating other types of medical data, such as patient history or related lab test results, into the automated report generation process. This could be achieved through advanced data fusion techniques, enhancing both the scope and reliability of the reports.
- **Interpretable Models:** As machine learning models become increasingly complex, the need for interpretability grows. Developing methods that offer insights into the model's decision-making process could not only improve trust in automated systems but also yield new clinical insights.

- **Real-World Deployment:** Lastly, turning these theoretical advancements into practical tools requires consideration of deployment challenges, such as data privacy concerns, real-time processing requirements, and integration with existing healthcare IT systems.

Given the rapidly evolving landscape of both medical imaging technology and machine learning methodologies, the field of automated radiology report generation is primed for continued growth and innovation. These future directions not only offer avenues for academic exploration but also hold the potential to significantly impact clinical practice, thereby augmenting the capabilities and efficiencies of medical diagnostics.

References

- [Alexander *et al.*2020] Robert G Alexander, Stephen Waite, Stephen L Macknik, and Susana Martinez-Conde. 2020. What do radiologists look for? advances and limitations of perceptual learning in radiologic search. *Journal of Vision*, 20(10):17–17.
- [Anderson *et al.*2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society.
- [Ba *et al.*2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- [Banerjee and Lavie2005] Satanjeev Banerjee and Alon Lavie. 2005. ME-TEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.
- [Bowman *et al.*2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.
- [Brown *et al.*2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

- [Chen *et al.*2020a] Y. C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. 2020a. Uniter: Universal image-text representation learning. In *European conference on computer vision*.
- [Chen *et al.*2020b] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020b. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1439–1449. Association for Computational Linguistics.
- [Chen *et al.*2021] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5904–5914. Association for Computational Linguistics.
- [Chowdhery *et al.*2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- [Cornia *et al.*2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. Computer Vision Foundation / IEEE.
- [Demner-Fushman *et al.*2016] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc.*, 23(2):304–310.
- [Devlin *et al.*2015] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *Computer Science*.
- [Devlin *et al.*2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume*

- l (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- [Donahue *et al.*2015] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2625–2634. IEEE Computer Society.
- [Dosovitskiy *et al.*2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- [Faghri *et al.*2018] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.
- [Fang *et al.*2022] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17988–17998. IEEE.
- [Hartung *et al.*2020] Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. 2020. How to create a great radiology report. *RadioGraphics*, 40(6):1658–1670.
- [He *et al.*2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- [He *et al.*2020] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*.
- [Herdade *et al.*2019] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Conference and Workshop on Neural Information Processing Systems*.

- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- [Hosny *et al.*2018] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.
- [Hu *et al.*2018] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141. Computer Vision Foundation / IEEE Computer Society.
- [Hu *et al.*2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- [Huang *et al.*2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- [Huang *et al.*2019a] Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019a. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4633–4642. IEEE.
- [Huang *et al.*2019b] Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. 2019b. Multi-attention and incorporating background information model for chest x-ray image report generation.
- [Huang *et al.*2023] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818.
- [Irvin *et al.*2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.

- [Jain *et al.*2021] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre J. Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- [Jin *et al.*2023] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2023. Promptmrg: Diagnosis-driven prompts for medical report generation. *arXiv preprint arXiv:2308.12604*.
- [Jing *et al.*2018] Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2577–2586. Association for Computational Linguistics.
- [Jing *et al.*2019] Baoyu Jing, Zeya Wang, and Eric P. Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6570–6580. Association for Computational Linguistics.
- [Johnson *et al.*2019] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.
- [Karpathy and Fei-Fei2015] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.
- [Katzman *et al.*2023] Benjamin D Katzman, Christian B van der Pol, Philippe Soyer, and Michael N Patlas. 2023. Artificial intelligence in emergency radiology: A review of applications and possibilities. *Diagnostic and Interventional Imaging*, 104(1):6–10.
- [Kim *et al.*2017] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference*

Track Proceedings. OpenReview.net.

- [Kingma and Ba2015] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- [Kiros *et al.*2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*.
- [Kisilev *et al.*2015a] P. Kisilev, E. Walach, E. Barkan, B. Ophir, and S. Y. Hashoul. 2015a. From medical image to automatic medical report generation. *Ibm Journal of Research and Development*, 59(2/3):2:1–2:7.
- [Kisilev *et al.*2015b] P. Kisilev, E. Walach, S. Hashoul, E. Barkan, and S. Alpert. 2015b. Semantic description of medical image findings: structured learning approach. In *British Machine Vision Conference 2015*.
- [Kisilev *et al.*2016] P. Kisilev, E. Sason, E. Barkan, and S. Hashoul. 2016. Medical image description using multi-task-loss cnn. In *International conference on medical image computing and computer assisted intervention; Deep learning in medical image analysis workshop; Workshop on large-scale annotation of biomedical data and expert label synthesis*, volume 10008, pages 121–129.
- [Kojima *et al.*2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- [Krause *et al.*2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345. IEEE Computer Society.
- [Lee *et al.*2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 212–228. Springer.
- [Li *et al.*2018] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen

- Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.
- [Li *et al.*2019a] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019a. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Hawaii, USA, January 27 - February 1, 2019*, pages 6666–6673. AAAI Press.
- [Li *et al.*2019b] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019b. Entangled transformer for image captioning. In *IEEE International Conference on Computer Vision*.
- [Li *et al.*2021] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. 2021. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3799–3802.
- [Li *et al.*2023a] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.
- [Li *et al.*2023b] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023b. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web (WWW)*, 26(1):253–270.
- [Lin *et al.*2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- [Lin *et al.*2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1449–1457. IEEE Computer Society.
- [Lin2004] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

- [Liu *et al.*2017] Siqu Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 873–881. IEEE Computer Society.
- [Liu *et al.*2019] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 3–11. ACM.
- [Liu *et al.*2020] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- [Liu *et al.*2021a] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13753–13762. Computer Vision Foundation / IEEE.
- [Liu *et al.*2021b] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*.
- [Liu *et al.*2021c] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- [Lu *et al.*2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society.
- [Ouyang *et al.*2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*,

35:27730–27744.

- [Pan *et al.*2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10968–10977. Computer Vision Foundation / IEEE.
- [Papineni *et al.*2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.
- [Pasunuru and Bansal2017] Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 979–985. Association for Computational Linguistics.
- [Paszke *et al.*2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.
- [Pavlopoulos *et al.*2022] John Pavlopoulos, Vasiliki Kougia, Ion Androutsopoulos, and Dimitris Papamichail. 2022. Diagnostic captioning: a survey. *Knowl. Inf. Syst.*, 64(7):1691–1722.
- [Qin *et al.*2019] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. 2019. Look back and predict forward in image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8367–8375. Computer Vision Foundation / IEEE.
- [Radford *et al.*2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

- [Radford *et al.*2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- [Raffel *et al.*2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.
- [Reimers and Gurevych2019] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- [Ren *et al.*2017a] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017a. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.
- [Ren *et al.*2017b] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017b. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.
- [Rennie *et al.*2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.
- [Scao *et al.*2022] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- [Schroff *et al.*2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

- [Sim *et al.*2023] Jordan ZT Sim, KN Bhanu Prakash, Wei Min Huang, and Cher Heng Tan. 2023. Harnessing artificial intelligence in radiology to augment population health. *Frontiers in Medical Technology*, 5.
- [Su *et al.*2020] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- [Tanida *et al.*2023] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.
- [Tay *et al.*2023] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):109:1–109:28.
- [Touvron *et al.*2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- [Vaswani *et al.*2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- [Vedantam *et al.*2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.
- [Vinyals *et al.*2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

- [Waite *et al.*2019] Stephen Waite, Arkadij Grigorian, Robert G Alexander, Stephen L Macknik, Marisa Carrasco, David J Heeger, and Susana Martinez-Conde. 2019. Analysis of perceptual expertise in radiology—current knowledge and a new perspective. *Frontiers in human neuroscience*, 13:213.
- [Wang *et al.*2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5005–5013. IEEE Computer Society.
- [Wang *et al.*2020] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, pages 18–34. Springer.
- [Wang *et al.*2021] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. 2021. A self-boosting framework for automated radiographic report generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2433–2442. Computer Vision Foundation / IEEE.
- [Wang *et al.*2022a] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2022a. On distinctive image captioning via comparing and reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [Wang *et al.*2022b] Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. 2022b. Automated radiographic report generation purely on transformer: A multi-criteria supervised approach. *IEEE Transactions on Medical Imaging*.
- [Wang *et al.*2022c] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022c. A medical semantic-assisted transformer for radiographic report generation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part III*, volume 13433 of *Lecture Notes in Computer Science*, pages 655–664. Springer.
- [Wang *et al.*2023] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11558–11567. IEEE.

- [Wei *et al.*2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- [Williams *et al.*2018] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- [Wolf *et al.*2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.
- [Wu *et al.*2020] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. 2020. Fine-grained image captioning with global-local discriminative objective. *IEEE Trans. on Multimedia*.
- [Xu *et al.*2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.
- [Xue and Huang2019] Yuan Xue and Xiaolei Huang. 2019. Improved disease classification in chest x-rays with transferred features from report generation. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging - 26th International Conference, IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings*, volume 11492 of *Lecture Notes in Computer Science*, pages 125–138. Springer.

- [Xue *et al.*2018] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer K. Antani, George R. Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, volume 11070 of *Lecture Notes in Computer Science*, pages 457–466. Springer.
- [Yang *et al.*2016] Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Review networks for caption generation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2361–2369.
- [Yang *et al.*2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10685–10694. Computer Vision Foundation / IEEE.
- [Yang *et al.*2020a] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. 2020a. Covid-ct-dataset: A ct scan dataset about covid-19.
- [Yang *et al.*2020b] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2020b. Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [Yang *et al.*2021a] S. Yang, X. Wu, S. Ge, X. Wu, S. K. Zhou, and L. Xiao. 2021a. Radiology report generation with a learned knowledge base and multi-modal alignment. *Image and Video Processing*.
- [Yang *et al.*2021b] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao. 2021b. Knowledge matters: Radiology report generation with general and specific knowledge. *Medical Image Analysis*.
- [Yao *et al.*2018] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 711–727. Springer.

- [Yin *et al.*2019] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In Jianyong Wang, Kyuseok Shim, and Xindong Wu, editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 728–737. IEEE.
- [You *et al.*2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4651–4659. IEEE Computer Society.
- [You *et al.*2021] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III*, volume 12903 of *Lecture Notes in Computer Science*, pages 72–82. Springer.
- [Yu *et al.*2019] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.
- [Yuan *et al.*2019] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI*, volume 11769 of *Lecture Notes in Computer Science*, pages 721–729. Springer.
- [Zhang *et al.*2020] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [Zhang *et al.*2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

- [Zhang *et al.*2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- [Zhou *et al.*2020] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4776–4785. Computer Vision Foundation / IEEE.