Chest Radiology Report Generation

Wei-Han Hsu wh2405@nyu.edu
Center for Data Science
New York University

Editor: Wei-Han Hsu

Abstract

Automatic generation of chest radiology reports is crucial for handling the growing volume of chest X-ray images. This study investigates a novel approach using transfer learning to adapt a state-of-the-art model (based on BLIP2 architecture) to a new chest X-ray dataset (Indiana University dataset). The proposed model incorporates both visual features extracted from chest X-rays and structured pathology findings. By leveraging parameter-efficient fine-tuning, the model effectively integrates this information with a Large Language Model (LLM) for report generation, while simultaneously adapting to the IU chest X-ray data. This approach holds promise for alleviating the workload of radiologists and improving healthcare efficiency.
Repository: https://github.com/zoe70416/chest-xray-report-generation

Keywords: BLIP2, Large Language Model, Vision Transformer, Image Encoder, Transfer Learning, Prompt Engineering, MIMIC-CXR, Indiana University Chest Xray

1 Purpose

This project aims to develop a generalizable chest X-ray report generation model by leveraging transfer learning from state-of-the-art multimodal approaches. Our model will integrate visual features from chest X-rays with structured pathology findings from Chexpert. By adapting to diverse datasets, this approach has the potential to enhance diagnostic capabilities and improve healthcare efficiency.

2 Introduction

Medical imaging serves as a cornerstone in modern diagnostics, with X-rays playing a pivotal role due to their affordability and widespread availability in hospitals worldwide. They are particularly invaluable in analyzing lung conditions, aiding in the diagnosis and monitoring of diseases such as pneumonia, pneumothorax, atelectasis, and edema (Johnson et al. (2019)). However, the sheer volume of chest X-rays requiring examination in daily clinical practice poses a significant challenge, compounded by a shortage of trained radiologists in many healthcare systems (Rosenkrantz AB and Jr (2016)). Consequently, automatic radiology report generation has emerged as a promising avenue of research to alleviate radiologists' workload.

Generating radiology reports presents complexities, as they entail multiple sentences, each

describing specific medical observations in a particular anatomical region (Hou et al. (2021)). Current methods often yield factually incomplete and inconsistent reports, lacking key observations and containing erroneous information (Miura et al. (2021)). To address these challenges, this study harnesses pre-trained large-scale Vision-Language models and pathology classifiers to ensure that image features or reports capture essential observations accurately.

Viewed as an 'image captioning' task, radiology report generation necessitates exploration of multi-modal approaches capable of integrating image and text information effectively (Hossain et al. (2018), Selivanov (2023)). Large-scale Vision-Language models, at the forefront of exploration in computer vision and natural language processing, enable machines to comprehend and generate information from visual and textual inputs. These models bridge the gap between visual understanding and language comprehension, demonstrating remarkable capabilities in tasks such as image captioning , visual question answering, and visual commonsense reasoning. Leveraging extensive image and text datasets, along with fine-tuning using task-specific data, enhances model performance and alignment with specific end goals and user preferences.

Building upon Chantal's pioneering work (Pellegrini et al. (2023), this study adopts the BLIP-2 (Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models) architecture to integrate image features and structured pathology findings with a Large Language Model (LLM) (Li et al. (2023)). While the previous study utilized the MIMIC dataset for training, this work further refines the model performance by fine-tuning with LoRA and transferring learning to the Indiana University Chest X-ray dataset.

3 Hypothesis

Our study assumed a high degree of similarity between the image feature and image-text spaces of the MIMIC and IU chest X-ray datasets. This assumption allowed us to explore fine-tuning the model on the IU dataset.

4 Data

This study utilizes two publicly available chest X-ray datasets: MIMIC-CXR and IU X-ray.

1. MIMIC-CXR: This dataset provides 227,827 free-text radiology reports along with corresponding chest X-ray images in JPG format (MIMIC-CXR-JPG v2.0.0, 377,110 images). MIMIC-CXR offers pre-defined train/validation/test splits for the reports. Notably, the image dataset (MIMIC-CXR-JPG) includes various view positions like AP (anterior-posterior), PA (postero-anterior), and lateral.

2. IU X-ray: This dataset consists of 7,470 chest X-ray images with paired radiology reports (3,955 reports). The IU X-ray dataset utilizes a patient-based split for training, validation, and testing data (60-20-20).

5 Material and Methods

5.1 Material

This study employed several pre-trained models from previous studies, each serving a specific purpose:

(1) *BioViL-T Image Encoder*: BioViL-T is a domain-specific X-ray encoder that utilizes pre-trained weights from ResNet-50. This frozen image encoder processes X-ray images, extracting relevant visual features to its' pair chest X-ray report. It was trained on the MIMIC-CXR dataset.(Bannur et al. (2023))

(2) *Chexbert*: This model, provided with pre-trained weights, serves as an accurate deep-learning-based chest radiology report labeler. It labels 14 medical observations, including: Fracture, Consolidation, Enlarged Cardiomediastinum, No Finding, Pleural Other, Cardiomegaly, Pneumothorax, Atelectasis, Support Devices, Edema, Pleural Effusion, Lung Lesion, Lung Opacity, facilitating comprehensive report generation. (Smit et al. (2020))

(3) *CheXpert Classifier*: Pretrained weights are utilized in this model to provide structured findings based on image features extracted by the BioViL-T Image Encoder. The model, trained separately using CheXbert labels from MIMIC-CXR reports, ensures the clinical efficacy of our approach (Irvin et al. (2019)).

(4) *Image Alignment Module*: Q-Former, a lightweight transformer, extracts relevant visual features from chest X-rays paired with radiology reports. Serving as an information bottleneck between the frozen image encoder and the frozen Large Language Model (LLM), it selects and feeds the most pertinent visual feature to the LLM for text generation. Patch-based features are transformed into 32 embedded language model tokens, ensuring accurate alignment with textual prompts.

(5) *Prompt Construction*: This module converts image features, structured findings, and instructions into a single prompt, serving as input for the LLM. The exact formulations of instruction prompts are detailed in the supplementary material.

(6) *Large Language Model*: Utilizing vicuna - 7b, an LLM, this model processes the constructed prompt and generates context-specific responses, completing the radiology report generation process. (Zheng et al. (2023))

5.2 Model Architecture

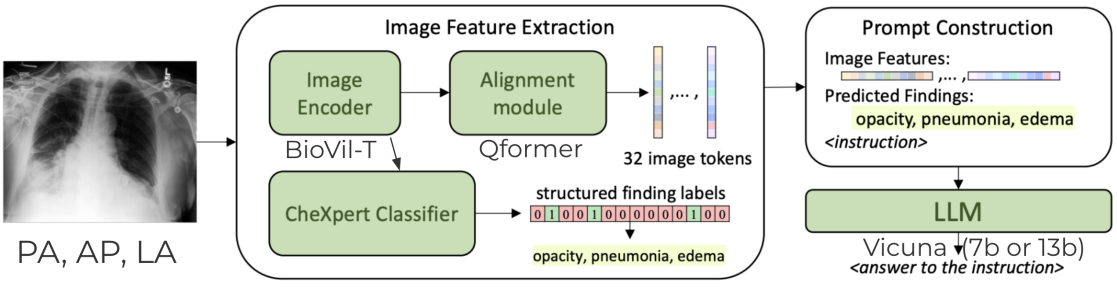

Figure 1: The model architecture of the proposed model

5.3 Methods

Both the alignment model and the LLM underwent further training/fine-tuning using the IU dataset.

(1) *Alignment Model*: Building on BLIP-2's stage-1 pre-training (Image-Text Contrastive Learning, Image-Grounded Text Generation, Image-Text Matching), we fine-tuned the alignment model with visual language training on IU X-ray image-report pairs (Li et al. (2023)). Training loss was monitored to prevent overfitting, as illustrated in Figure 2. The model checkpoint at epoch 13, exhibiting the lowest training loss, was selected.

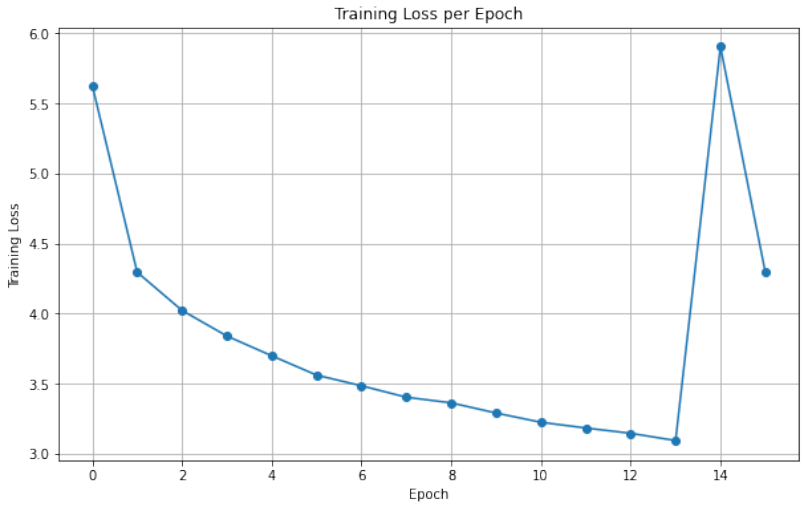

Figure 2: The training loss during fine-tuning/training of the Alignment model

(2) *LLM Fine-tuning with LoRA*: LoRA, or Low-Rank Adaptation, optimizes large models by freezing pre-trained weights and introducing trainable rank decomposition matrices into each Transformer layer. This approach significantly reduces the number of trainable parameters for downstream tasks. The LLM (vicuna model) was fine-tuned using LoRA. The PEFT library from Hugging Face was employed, with a learning rate (LR) of 3×104, and training was conducted for up to five epochs, employing early stopping based on the validation set.

(3) *Prompt Engineer*: In addition to the basic prompt utilized in the prior study, three additional prompts were developed to target distinct aspects of the report generation process, potentially enhancing report quality:
- Focus on Specific Findings
- Focus on Chest X-ray Analysis
- Focus on Radiology Style

6 Results

To assess our model's performance, we employed Natural Language Generation (NLG) metrics:

(1) BLEU (Bilingual Evaluation Understudy): BLEU-1, BLEU-2, BLEU-3, BLEU-4 (B-4). This metric, ranging from zero to one, gauges the similarity between machine-translated text and a set of high-quality reference translations.

(2) METEOR (Metric for Evaluation of Translation with Explicit ORdering): METEOR compares the translated text to human reference translations by segmenting them into chunks and computing similarity using measures such as unigram precision, recall, F-score, bigram overlap, and exact word matches.

(3) ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE focuses on recall, determining how much information present in reference text(s) is captured by the generated text.

Figure 3 presents a table illustrating model performance. The initial row displays results from the preceding study, where the model underwent training and evaluation on the MIMIC-CXR dataset.

Subsequent rows showcase the performance of models evaluated on the IU dataset. The baseline model, from the prior study, was tested without additional tuning, as indicated in the second row (highlighted in yellow).

Both the alignment model and the LLM underwent further fine-tuning or training on the IU dataset, with specific prompts denoted by numbers. If no prompt was specified, the basic prompt was employed.

While our fine-tuned model achieved positive results on the IU dataset, the baseline model outperformed it on most evaluation metrics for chest X-ray reports. However, our model showed promise with the "Focus on Specific Findings" prompt, achieving a higher METEOR score.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
MIMIC-CXR_pretrained	0.340	-	-	0.097	0.136	0.270
IU_pretrained	**0.343**	**0.211**	**0.140**	**0.089**	0.141	**0.245**
IU_retrained (allign)	0.325	0.181	0.112	0.070	0.143	0.234
IU_retrained (allign+llm)	0.317	0.174	0.107	0.068	0.133	0.242
IU_retrained (allign+llm)-1	0.233	0.123	0.068	0.038	**0.157**	0.212
IU_retrained (allign+llm)-2	0.288	0.158	0.090	0.055	0.150	0.238
IU_retrained (allign+llm)-3	0.321	0.183	0.112	0.068	0.147	0.226

Figure 3: Performance comparisons of the proposed model

7 Discussions

This study explored the transferability of a BLIP2-based model for chest X-ray report generation on a new dataset (IU X-ray) using transfer learning. We hypothesized that a high degree of similarity between the image feature and image-text spaces of the MIMIC

and IU datasets would enable successful fine-tuning of the model. While the fine-tuned model achieved positive results on the IU dataset as measured by NLG metrics (BLEU, METEOR, ROUGE), the baseline model, directly transferred from MIMIC, outperformed it on most metrics.

These findings suggest that potential discrepancies exist between the image feature and image-text spaces of the two datasets. Here, we propose several future research directions to address these limitations and improve the model's generalizability

Future research directions include:

1. **Addressing Dataset Disparity**: Explore data augmentation techniques to expand the IU dataset, potentially bridging the gap with MIMIC. This involves generating synthetic chest X-rays or modifying existing ones to increase dataset size and enhance model performance. Additionally, investigate domain adaptation methods to help the model adapt to the specific characteristics of the IU dataset, bridging the gap between the MIMIC and IU data distributions for effective learning.

2. **Advanced Fine-tuning Techniques**: Exploring alternative fine-tuning approaches beyond LoRA might enhance the model's adaptation to the IU dataset. Also, training the LLM for more epochs could potentially improve its performance, but careful monitoring for overfitting is crucial.

3. **Investigate Larger LLMs**: Consider exploring the use of larger LLM architectures, such as Vicuna-13b, for chest X-ray report generation. This could potentially lead to improvements in report quality and detail.

4. **Model Evaluation with Domain Experts**: Including radiologists in the evaluation process could provide valuable insights beyond NLG metrics, ensuring the generated reports are clinically relevant and accurate.

This study explored the transferability of a BLIP2-based model for chest X-ray report generation on a new dataset. While the fine-tuned model achieved positive results, room for improvement exists. By addressing the limitations and exploring future directions, this research paves the way for developing robust and generalizable models for automatic radiology report generation.

8 Student Contributions

This project was solely undertaken by Wei-Han Hsu.

Acknowledgments and Disclosure of Funding

My sincere gratitude to Dr. Razavian, and Dr. Deniz both from the Department of Radiology at NYU Grossman School of Medicine), and the NYUIT HPC staff for their guidance and support.

Appendix A.

In this appendix, we'll provide the prompt we used to generate the report :

Basic prompt "Given the provided image ¡IMG¿ and predicted findings (findings), generate a concise chest X-ray report in the style of a radiologist. Prioritize these findings while maintaining clarity and conciseness."

Focus on Specific Findings "Image information: ¡IMG¿. Predicted Findings: findings. Focus on these findings and write a concise chest X-ray report in the style of a radiologist. Emphasize the reported findings while keeping the report clear and informative."

Chest X-ray Analysis "Given the provided image ¡IMG¿ and predicted findings (findings), generate a concise chest X-ray report in the style of a radiologist. Prioritize these findings while maintaining clarity and conciseness."

Focus on Radiology Style "Given the image data ¡IMG¿ and predicted findings (findings), create a chest X-ray report that emulates the conciseness and clarity of a radiologist's report. Emphasize the identified findings while keeping the report informative."

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023.

Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning, 2018.

Daibing Hou, Zijian Zhao, Yuying Liu, Faliang Chang, and Sanyuan Hu. Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, 9: 21236–21250, 2021. doi: 10.1109/ACCESS.2021.3056175.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie 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. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance, 2023.

Hughes DR Rosenkrantz AB and Duszak R Jr. The u.s. radiologist workforce: An analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279,1: 175–84, 2016. doi: 10.1148/radiol.2015150921.

Rogov O.Y. Chesakov D. Selivanov, A. Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13,1:4171, 2023. doi: 10.1038/s41598-023-31223-5.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.