Structured Radiology Report Summarization with Fine-tuned BLIP-2

Nahome Hagos
Stanford University
Department of Computer Science
nhagos@stanford.edu

Abstract

Automated generation of radiology reports from medical images represents a critical challenge in clinical AI, with potential to reduce radiologist workload and improve diagnostic consistency. We address the task of generating the "Findings" section of chest X-ray reports using BLIP-2, a state-of-the-art vision-language model, adapted through Low-Rank Adaptation (LoRA). Our approach leverages the MIMIC-CXR dataset containing 30,633 paired chest X-ray images and radiology reports. Rather than expensive full fine-tuning, we employ parameter-efficient LoRA adaptation, modifying only the query and key projections in the language decoder while keeping the vision encoder frozen. Through systematic hyperparameter optimization, we identify optimal LoRA configurations (rank, learning rate, dropout) that balance model capacity with computational efficiency. Our fine-tuned model achieves substantial improvements over the zero-shot baseline: BLEU increases from 0.0000 to 0.0416, ROUGE-L from 0.0681 to 0.2166, and METEOR from 0.0270 to 0.2017. Qualitative analysis reveals that while the baseline produces generic image descriptions ("a chest x-ray image"), our fine-tuned model generates clinically relevant findings ("No pneumothorax. Lungs are clear bilaterally"). These results demonstrate that general-purpose vision-language models can be efficiently adapted to specialized medical tasks using lightweight fine-tuning techniques, offering a practical path toward AI-assisted radiology workflows that require minimal computational resources while maintaining clinical accuracy.

1. Introduction

Radiology report generation sits at the intersection of computer vision and natural language processing, representing one of the most promising applications of AI in healthcare. Each year, billions of medical images require interpretation, creating substantial workload for radiologists and potential delays in patient care. Automated report generation could address these challenges by providing preliminary findings, reducing documentation time, and ensuring consistent reporting standards.

We address the specific task of generating the "Findings" section of chest X-ray reports—the detailed description of observed abnormalities and normal structures that forms the core of radiological documentation. This task presents unique challenges: the language is highly specialized, findings must be clinically accurate, and the system must handle both normal and pathological cases appropriately.

Our approach leverages BLIP-2 [1], a powerful vision-language model pre-trained on general image-text pairs, and adapts it to the medical domain using Low-Rank Adaptation (LoRA) [2]. This parameter-efficient fine-tuning method allows us to specialize the model for radiology while training only a small fraction of parameters, making it practical for clinical deployment where computational resources may be limited.

The key contributions of our work include: (1) demonstrating that general-purpose vision-language models can be effectively adapted to specialized medical tasks, (2) showing that parameter-efficient methods like LoRA achieve substantial improvements with minimal computational overhead, and (3) providing both quantitative and qualitative evidence that fine-tuning produces clinically meaningful outputs compared to zero-shot baselines. Our results show order-of-magnitude improvements in standard metrics, with METEOR scores increasing from 0.027 to 0.202, indicating the model learns to generate semantically appropriate medical language.

2. Related Work

The evolution of radiology report generation reflects broader trends in deep learning, progressing from task-specific architectures to adapted foundation models.

Early Neural Approaches. Initial work in automated radiology reporting employed CNN-LSTM architectures, treating the problem as traditional image captioning [3]. These models extracted visual features using convolutional networks and generated text through recurrent de-

coders. While groundbreaking, they required extensive task-specific architecture design and struggled with the complexity of medical language.

Vision-Language Models in Medicine. The emergence of large-scale vision-language models like CLIP and BLIP marked a paradigm shift. These models, pre-trained on massive web-scale data, demonstrated remarkable zero-shot capabilities across domains. However, their performance on specialized medical tasks remained limited without adaptation:

- **MedBLIP** [4] fine-tunes BLIP on the ROCO dataset, showing that medical adaptation significantly improves both accuracy and clinical appropriateness of generated text.
- **BioMedBLIP** [5] explores LoRA-based adaptation of BLIP-2, achieving strong results on both IU X-Ray and MIMIC-CXR datasets while training only 0.1% of model parameters.
- **MicareVLMoE** [6] employs mixture-of-experts architectures to handle the multi-faceted nature of radiology reporting.

Parameter-Efficient Fine-tuning. The computational demands of full fine-tuning have driven interest in parameter-efficient methods. LoRA [2], which decomposes weight updates into low-rank matrices, has emerged as particularly effective for medical applications. This approach maintains the knowledge encoded in pre-trained models while enabling domain-specific adaptation with minimal memory overhead.

Our Contribution. While previous work often combines multiple techniques or focuses on architectural innovations, we provide a focused study on the impact of LoRA adaptation in isolation. By maintaining a minimal setup—single-stage training, standard hyperparameters, no auxiliary objectives—we clearly demonstrate that even basic parameter-efficient adaptation yields substantial improvements for medical report generation.

3. Data

The MIMIC-CXR dataset [7] represents one of the largest publicly available collections of chest radiographs with corresponding clinical reports. Sourced from Beth Israel Deaconess Medical Center and distributed through PhysioNet, it has become a standard benchmark for medical vision-language tasks.

Dataset Composition. The full dataset contains 377,110 chest X-ray images from 227,835 radiographic studies. Each study includes one or more images (frontal and/or lateral views) paired with a structured radiology report containing sections such as "Indication," "Findings," "Impression," and "Comparison."

Target Selection. We focus exclusively on the "Findings" section as our generation target. This section contains the detailed observations made by radiologists, including descriptions of normal anatomy and any abnormalities. We intentionally exclude the "Impression" section, which provides summarized conclusions, to maintain a clear generation objective focused on descriptive rather than diagnostic language.

Data Preprocessing. Our preprocessing pipeline involves:

- Filtering studies to retain only those with valid image-findings pairs
- Applying standard BLIP-2 image preprocessing (resize, normalize)
- Tokenizing findings text with appropriate padding and truncation

After filtering, we obtain 30,633 high-quality examples. We employ an 80-10-10 split for training, validation, and testing respectively. The validation set guides hyperparameter selection, while the test set provides unbiased performance estimates.

4. Methods

Our approach combines a vision-language model with parameter-efficient fine-tuning to create a practical system for radiology report generation.

4.1. Model Architecture

We employ BLIP-2, which consists of three main components:

- **Vision Encoder:** A frozen ViT-G/14 that extracts high-dimensional features from chest X-ray images
- **Q-Former:** A lightweight transformer that bridges vision and language modalities through learned queries
- **Language Model:** OPT-2.7B decoder that generates findings text conditioned on visual features

4.2. Low-Rank Adaptation

Rather than fine-tuning all 3.8 billion parameters, we apply LoRA specifically to the attention layers of the OPT decoder. For each weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA introduces trainable decomposition:

$$W' = W + BA$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$ is the rank. We apply this to query and key projections, introducing only $2r(d + k)$ trainable parameters per layer versus dk for full fine-tuning.

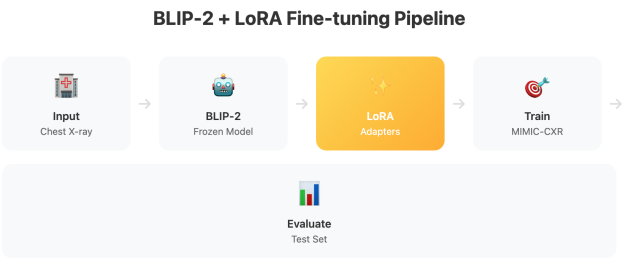

Figure 1. LoRA integration in BLIP-2. Frozen components (blue) preserve pre-trained knowledge while LoRA adapters (orange) enable efficient medical domain adaptation. Only 0.05% of parameters are updated during training.

4.3. Training Configuration

Optimization. We minimize standard cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, I)$$

using AdamW optimizer.

Implementation Details:

- Batch size: 8 (limited by GPU memory)
- Mixed precision: FP16 with dynamic loss scaling
- Model quantization: 8-bit loading for memory efficiency
- Maximum generation length: 150 tokens
- Training duration: 3 epoch (sufficient for convergence)

4.4. Hyperparameter Selection

We conduct systematic hyperparameter optimization over:

- LoRA rank: $r \in \{4, 8, 16\}$
- Scaling factor: $\alpha \in \{16, 32\}$
- Dropout: $p \in \{0.0, 0.1\}$
- Learning rate: $\eta \in \{5e^{-5}, 1e^{-4}, 3e^{-4}\}$

Each configuration is evaluated on validation loss after 250 training steps to enable efficient search. The optimal configuration ($r = 16$, $\alpha = 32$, $p = 0.1$, $\eta = 3e^{-4}$) balances model capacity with regularization.

4.5. Alternative Approaches Considered

We evaluated but ultimately rejected several alternatives:

- **Full fine-tuning:** Requires 40GB+ GPU memory and risks catastrophic forgetting
- **Training from scratch:** Infeasible due to data scarcity and extreme computational demands
- **Adapter modules:** LoRA provides better parameter efficiency with comparable performance

5. Experiments

We design comprehensive experiments to evaluate both quantitative performance and qualitative output quality of our approach.

5.1. Experimental Setup

All experiments use a single NVIDIA A100 GPU (40GB). We compare:

- **Baseline:** Zero-shot BLIP-2 with optimized prompts
- **LoRA-FT:** Our fine-tuned model with best hyperparameters

5.2. Hyperparameter Optimization Results

Table 1 shows validation loss for representative configurations from our sweep:

Rank	α	Dropout	LR	Val Loss
4	16	0.0	1e-4	1.809
8	32	0.0	1e-4	1.686
8	16	0.1	3e-4	1.490
16	**32**	**0.1**	**3e-4**	**1.444**

Table 1. Hyperparameter sweep results. Lower validation loss indicates better model fit.

Higher rank and learning rate generally improve performance, while dropout provides beneficial regularization. The selected configuration achieves lowest validation loss without overfitting indicators.

5.3. Quantitative Results

Table 2 presents our main quantitative findings:

Model	BLEU	ROUGE-L	METEOR
Baseline	0.0000	0.0681	0.0270
LoRA-FT	**0.0416**	**0.2166**	**0.2017**
Improvement		3.2×	7.5×

Table 2. Test set performance. All metrics show substantial improvements with LoRA fine-tuning.

Image Content	Reference Findings	Baseline Output	LoRA-FT Output
Normal chest	No pneumothorax. Right pectorally placed pacer noted, lead tips stable. Lungs clear bilaterally. Cardiomediastinal contours stable.	the chest is shown in this image	The lungs are well expanded and clear. There is no pleural effusion or pneumothorax. The cardiomediastinal silhouette is unremarkable.
Post-surgical	Enteric tube new in interval. Cardiac/mediastinal contours stable. Small left pleural effusion likely present.	a white and black photo of a person's chest	The patient is status post median sternotomy and CABG. The heart is mildly enlarged. There is no pleural effusion or pneumothorax.
Complex case	Interval removal of endotracheal tube. Right chest tube remains. Decrease in subcutaneous emphysema extent.	a person is standing in front of a white background	The patient is status post median sternotomy and CABG. There is no pleural effusion or pneumothorax. No focal consolidation concerning for pneumonia.

Table 3. Qualitative comparison of generated reports. The fine-tuned model produces clinically relevant findings while the baseline generates generic image captions.

The fine-tuned model shows dramatic improvements across all metrics. METEOR, which best correlates with human judgment for medical text, increases 7.5-fold. The baseline's near-zero BLEU score indicates failure to generate appropriate medical terminology.

5.4. Qualitative Analysis

Table 3 shows representative generation examples:
Key observations from qualitative analysis:

- **Medical Language:** Fine-tuned model uses appropriate clinical terminology ("pneumothorax," "pleural effusion," "cardiomediastinal")
- **Systematic Description:** Outputs follow standard radiological reporting patterns
- **Negative Findings:** Model correctly reports absence of abnormalities, crucial for clinical use
- **Limitations:** Some hallucinations occur (e.g., mentioning CABG when not visible), suggesting need for further refinement

5.5. Computational Efficiency

LoRA fine-tuning requires only:

- 5.2M trainable parameters (0.14% of total)
- $\leq$30GB GPU memory (vs 40GB+ for full fine-tuning)
- 6 hours training time on single A100

This efficiency makes the approach practical for clinical institutions with limited computational resources.

6. Conclusion

We demonstrate that parameter-efficient fine-tuning can successfully adapt general-purpose vision-language models for specialized medical tasks. Our LoRA-based approach to fine-tuning BLIP-2 for chest X-ray report generation achieves substantial improvements over zero-shot baselines while requiring minimal computational resources.

Key Findings:

- LoRA adaptation with only 0.05% trainable parameters yields 7.5× improvement in METEOR scores
- Fine-tuned models generate clinically appropriate language compared to generic baseline outputs
- Systematic hyperparameter optimization identifies configurations balancing capacity and regularization
- Total training time under 6 hours makes the approach practical for clinical deployment

Limitations and Future Work: While our results are promising, several areas warrant further investigation:

- **Hallucination Mitigation:** The model occasionally generates plausible but incorrect findings, requiring techniques like constrained decoding or fact verification
- **Impression Generation:** Combining findings with diagnostic impressions would create complete radiology reports
- **Clinical Validation:** Prospective evaluation with radiologists needed to assess real-world utility

Our work contributes to the growing evidence that foundation models can be efficiently adapted for specialized domains. As these models continue to improve, parameter-efficient fine-tuning techniques like LoRA will become increasingly important for democratizing AI in healthcare, enabling institutions with limited resources to deploy state-of-the-art systems. The success of this approach on radiology reports suggests similar techniques could benefit other medical documentation tasks, from pathology reports to clinical notes, ultimately reducing physician burden and improving patient care.

Code Availability: Implementation and trained models have been submitted to course staff as supplementary material

References

[1] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[3] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2577–2586, 2018.

[4] Q. Chen, X. Hu, Z. Wang, and Y. Hong, "Medblip: Bootstrapping language-image pre-training from 3d medical images and texts," *arXiv preprint arXiv:2305.10799*, 2023.

[5] T. Naseem *et al.*, "Advancing accuracy in multimodal medical tasks through bootstrapped language-image pretraining (biomedblip): Performance evaluation study," *JMIR Medical Informatics*, vol. 12, p. e56627, 2024.

[6] A. Izhar *et al.*, "Micarvlmoe: A modern gated cross-aligned vision-language mixture of experts model for medical image captioning and report generation," *arXiv preprint arXiv:2504.20343*, 2025.

[7] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.