Reproducing Med-VQA: Visual Question Answering in Radiology

Aria Abrishamkar
University of Illinois Urbana-Champaign, Deep Learning for Healthcare (SP25)

Abstract

We reproduce the Med-VQA task proposed in Xu et al. (2023), which evaluates medical vision-language understanding through clinical visual question answering. Using the VQA-RAD dataset, we assess the performance of the BLIP-2 model under a zero-shot inference setting. We also attempt lightweight fine-tuning in Colab to explore domain adaptation. Despite significant implementation and runtime constraints, we document reproducibility challenges, insights from failure analysis, and discuss why domain-specific pretraining remains essential in clinical AI settings.

GitHub Repository: `https://github.com/Aria-007/DL4HC_Project/tree/main`
Note: No video presentation was submitted for this project.

1. Introduction

Visual Question Answering (VQA) tasks involve answering natural language questions about images and require joint reasoning over visual and textual information. In recent years, general-domain VQA models such as BLIP, Flamingo, and GPT-4V have shown impressive capabilities on open benchmarks like VQAv2, OK-VQA, and GQA.

However, the clinical domain presents unique challenges. Medical images often contain grayscale or low-contrast visual signals, complex anatomical structures, and require precise domain-specific reasoning. In this context, hallucination or overgeneralization by vision-language models can have serious consequences. Furthermore, medical questions often rely on background clinical knowledge absent from general-purpose datasets.

Xu et al. (2023) address these challenges by pretraining a model on clinical image-text pairs from MIMIC-CXR, creating a model better aligned with tasks like Med-VQA. In this project, we reproduce the Med-VQA setup by evaluating the performance of BLIP-2, a state-of-the-art generalist model, on the VQA-RAD dataset under a zero-shot setting. We also report on an attempted fine-tuning procedure and its failure due to resource limitations in Colab.

2. Dataset

VQA-RAD is a benchmark dataset for medical visual question answering, consisting of 3,515 QA pairs across 315 de-identified radiology images. The questions span five categories:

- **Modality:** e.g., ”What type of scan is shown?”
- **Plane/View:** e.g., ”Is this a coronal view?”
- **Organ:** e.g., ”Which organ is shown?”
- **Abnormality Presence:** e.g., ”Is there a mass in the lung?”
- **Reasoning:** e.g., ”What findings are indicated in this scan?”

Answers can be binary (“yes” or “no”), short phrases (“lungs”, “CT”), or complete sentences. Approximately 43% of questions are yes/no, while 57% are open-ended. We used a 50-sample subset due to Colab’s resource limitations.

3. Model

We used the BLIP-2 checkpoint `Salesforce/blip2-opt-2.7b` from Hugging Face, featuring:

- Vision Encoder (ViT-G)
- Q-Former attention module
- OPT-2.7B language decoder

This model is not trained on radiology data, which makes it a good test of out-of-domain generalization.

4. Implementation and Inference

The pipeline included image parsing, BLIP-2 tokenization, and generation in Google Colab (T4 GPU). We observed:

- 20+ minute image folder uploads
- File mismatches in JSON vs. folder
- 1–2 min inference time per sample
- Multiple crashes due to GPU RAM limits

5. Results and Error Analysis

Exact Match Accuracy: 0 / 50 = 0.00%

Table 1: *
Sample Prediction Errors

Q: What organ is shown?	
GT: lungs	**Pred:** heart
Q: Is there a mass in the right lung?	
GT: yes	**Pred:** no abnormality
Q: What modality is used?	
GT: CT	**Pred:** X-ray

Table 2: *
Error Type Breakdown

Error Type	Count
Modality Misclassification	15
Wrong Organ	12
Yes/No Incorrect	14
Descriptive Reasoning Failures	6
Image Not Found	3

6. Fine-Tuning Attempt

We tried to fine-tune only the language modeling head on 10 samples. Code snippet:

```
for param in model.parameters():
    param.requires_grad = False
for param in model.language_model.parameters():
    param.requires_grad = True

model.train()
optimizer = AdamW(model.language_model.parameters(), lr=2e-5)
```

All attempts failed with CUDA OOM errors, even with batch size 1 and cache clearing.

7. Discussion

These results confirm Xu et al.'s hypothesis: general-purpose VL models struggle in clinical domains. Despite BLIP-2's strong open-domain capabilities, its lack of medical vocabulary, report alignment, and domain supervision leads to poor performance and frequent hallucinations.

8. Future Work

- Use domain-specific VLMs (e.g., MedViLL, VLT)
- Try prompt tuning or LoRA on smaller models (e.g., BioGPT)
- Evaluate retrieval-augmented LLMs

9. Conclusion

We attempted to reproduce Med-VQA using BLIP-2 and VQA-RAD. The results were not successful, confirming the importance of clinical pretraining. Reproducibility in healthcare AI remains challenging under limited resources.

References

Xu, Y., Liu, P., Zhang, H., et al. (2023). *Multi-modal Pre-training for Medical Vision-language Understanding and Generation*. Proceedings of CHIL 2023.