UNIVERSITI MALAYA

A Comparative Study of CNN-Based and Vision-Language Models for Binary Medical Visual Question Answering

Name	Matric No.
Yousef H. A. Awad	S2122641
Faza Nanda Yudistira	S24061590

1. Background

1.1 Context

Medical imaging is a cornerstone of modern diagnostics, generating vast amounts of complex visual data such as X-rays, CT scans, and MRIs. Traditionally, interpreting these images requires highly trained radiologists to detect abnormalities and translate visual findings into clinical reports. With the advent of deep learning, Computer-Aided Diagnosis (CAD) systems-primarily based on Convolutional Neural Networks (CNNs)-have achieved remarkable success in classifying specific pathologies (He et al., 2016).

However, traditional image classification models operate as "black boxes" with rigid output classes. They lack the ability to interact with the clinician or answer specific, context-dependent inquiries. In a real-world clinical setting, a doctor is rarely interested in a simple class label; they often ask natural language questions such as, "Is the heart enlarged?", "Are there opacities in the lower left lobe?", or "Is the catheter positioned correctly?".

1.2 Problem

To bridge the gap between pixel-level information and natural language reasoning, Medical Visual Question Answering (Med-VQA) has emerged. Med-VQA systems take a medical image and a natural language question as input and generate a relevant answer.

While general-domain Vision-Language Models (VLMs) like GPT-4V or LLaVA (Liu et al., 2023) have shown impressive capabilities, their application in the medical domain remains challenging due to:

1. Data Scarcity: Medical VQA datasets are significantly smaller than general web-scraped datasets.
2. Domain Specificity: Medical terminology and subtle radiological features (e.g., small nodules) require specialized knowledge that general models often hallucinate or miss.
3. Compute Constraints: Fine-tuning massive VLMs requires significant resources, raising the question of whether smaller, specialized architectures might be more efficient for specific tasks.

This project addresses these challenges by conducting a comparative study between a traditional, lightweight Deep Learning baseline (CNN+RNN) and a modern, instruction-tuned VLM (LLaVA-Med) (Liu et al., 2023), specifically focusing on binary (Yes/No) clinical questions regarding Chest X-rays.

2. Objectives

The primary goal of this project is to evaluate the trade-offs between architectural complexity and performance in the context of binary Medical VQA. We aim to determine if a specialized, pre-trained biomedical VLM offers a statistically significant advantage over a standard multimodal baseline when restricted to a specific modality and question type.

2.1 Primary Objective

To compare the performance of a CNN-based Multimodal Baseline (ResNet + BiLSTM) against a state-of-the-art Vision-Language Model (LLaVA-Med) on the task of binary (Yes/No) Visual Question Answering for Chest X-rays.

2.2 Research Questions

To achieve the primary objective, this study aims to answer the following research questions:

1. Performance Comparison: Which approach achieves higher classification metrics (Accuracy, Macro-F1) on the binary subset of the VQA-RAD dataset?
2. Data Efficiency: Which approach is more robust when training data is severely limited? We hypothesize that the pre-trained knowledge of the VLM may allow it to perform better with fewer samples (Few-Shot/Low-Resource learning).
3. Error Analysis: What are the dominant failure modes for each architecture? Does the VLM struggle with negation (e.g., "Is there no evidence of...") or specific spatial reasoning compared to the baseline?

2.3 Success Metrics

As this is a binary classification task (answering "Yes" or "No"), we will utilize standard classification metrics:

- Accuracy: Overall correctness of predictions.
- Macro-F1 Score: To account for potential class imbalances in the filtered dataset.
- Data Efficiency Curve: A plot of performance (y-axis) vs. percentage of training data used (x-axis).

3. Method: Dataset, Preparation, and Preprocessing

3.1 Dataset Source: VQA-RAD

We utilize the VQA-RAD dataset (Lau et al., 2018), a clinician-generated radiology Visual Question Answering dataset released on the Open Science Framework (OSF). Unlike automatically generated datasets (like Slake or VQA-Med-2019), VQA-RAD consists of naturally occurring clinical questions validated by radiologists, ensuring high clinical relevance.

The raw dataset contains varying body parts (Head, Chest, Abdomen) and modalities (CT, MRI, X-ray). The total dataset size reported in the literature is 315 images and 3,515 QA pairs.

3.2 Dataset Filtering and Scope

To ensure a controlled experiment and reduce confounding variables (such as model difficulty in distinguishing MRI from X-ray), we apply a strict filtering pipeline:

1. Modality Restriction (Chest X-ray Only): We restrict the dataset to Chest X-rays. According to the dataset metadata, this subset consists of approximately 107 images. This focus allows us to evaluate the models on the most common radiological modality.
2. Answer Type Restriction (Binary Only): We further filter the samples to include only Closed-Ended questions where the ground-truth answer maps to {Yes, No}.
   - Justification: Open-ended generation poses evaluation challenges (e.g., synonym matching). Binary classification provides a rigorous, objective metric for comparing the reasoning capabilities of the two models.

3.3 Data Splitting Strategy (Leakage Prevention)

A critical issue in medical machine learning is data leakage, where slices from the same patient or augmentations of the same image appear in both training and test sets, artificially inflating accuracy.

To prevent this, we implement Image-Level Splitting:

- All questions associated with a specific image ID are grouped together.
- We split the unique Images into Train (80%), Validation (10%), and Test (10%) sets.
- This ensures the model is tested on unseen medical cases, not just unseen questions about familiar images.

3.4 Preprocessing Pipeline

We apply consistent preprocessing to ensure fair input conditions:

- Image Normalization:
    - CNN Baseline: Images are resized to 224x224 pixels and normalized using ImageNet (Russakovsky et al., 2015) mean/std.
    - LLaVA-Med: Images are resized to 336x336 pixels to match the CLIP (Radford et al., 2021) visual encoder requirements.
- Text Preprocessing:
    - Includes lowercasing and mapping variations like "YES" or "yes." to a single integer class {1:yes, 0:no}
- Augmentation (Training Only):
    - Text is standardized to binary integer classes. Conservative augmentations like rotation and brightness adjustments are applied, while flipping is avoided to preserve clinical laterality.

4. Method: Algorithms and Model Architectures

We compare two distinct deep learning paradigms: a standard modular baseline and an integrated Vision-Language Model.

4.1 Model A: CNN-Based Multimodal Baseline (The "Classic" Approach)

This model represents the standard deep learning approach to VQA prior to the Transformer era. It treats VQA as a classification task via feature fusion.

- Image Encoder (Vision): We use ResNet-50, pre-trained on ImageNet. We extract the global average pooled feature
- s from the final convolutional block. The weights can be frozen or fine-tuned.Question Encoder (Language): We use a Bi-Directional LSTM (BiLSTM). The questions are tokenized, embedded via GloVe or learned embeddings, and passed through the LSTM to capture sequential dependencies.
- Fusion Mechanism: The visual vector v and the question vector q are concatenated ($v \oplus q$) and passed through a Multi-Layer Perceptron (MLP).
- Output Layer: A final softmax layer with 2 neurons (Yes/No).
- Justification: This architecture is lightweight, easy to train, and serves as a strong baseline to determine if the complexity of a VLM is actually necessary for binary tasks.

4.2 Model B: Vision-Language Model - LLaVA-Med (The "SOTA" Approach)

We select LLaVA-Med (Large Language and Vision Assistant for BioMedicine), a specialized adaptation of the LLaVA architecture.

- Architecture: LLaVA-Med connects a visual encoder (CLIP ViT-L/14) to a Large Language Model (Vicuna/LLaMA) using a simple linear projection layer.
- Biomedical Alignment: Unlike standard GPT-4 or LLaVA, LLaVA-Med was fine-tuned on a large dataset of biomedical image-text pairs from PubMed Central, giving it domain-specific knowledge.
- Adaptation for Binary VQA:
  - We treat the task as text generation. We prompt the model: "You are a medical assistant. Look at this Chest X-ray and answer the question with only 'yes' or 'no'. Question: {Q}".
  - We employ Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Low-Rank Adaptation) (Hu et al., 2022). This allows us to fine-tune the model on our small VQA-RAD subset without requiring massive GPU memory, fitting the constraints of this academic project.
- Justification: LLaVA-Med represents the current state-of-the-art in open-source biomedical VQA. Using it tests the hypothesis that "general understanding" + "instruction following" is superior to specialized classification layers.

4.3 Training Protocol

To ensure a fair comparison:

- Both models use the exact same Train/Val/Test splits.
- Loss Function: Weighted Cross-Entropy (to handle any residual class imbalance).
- Stopping Criterion: Early stopping based on Validation Macro-F1 score to prevent overfitting on the small dataset.

5. Preliminary Results

5.1 Dataset Analysis and Distribution

We have successfully implemented the data ingestion and filtering pipeline. The breakdown of the VQA-RAD dataset for our specific scope is as follows:

1. Total Raw Samples: 3,515 QA pairs.
2. After Chest X-ray Filter: Reduced to ~107 unique images.

3. After Binary (Yes/No) Filter:
    - We retained approximately 450 QA pairs.
    - Class Balance: The distribution of answers is nearly balanced (approx. 50.5% "Yes" vs 49.5% "No").

Significance: The balanced nature of the subset confirms that accuracy will be a valid metric (we do not need to rely solely on F1-score due to severe imbalance). The sample size, while small, is consistent with "Few-Shot" benchmarks often used in medical imaging literature.

5.2 Qualitative Data Inspection

We inspected random samples to verify the data quality. The questions vary in complexity:

- Simple: "Is this a chest x-ray?" (Sanity check)
- Detection: "Is there a mass in the right lung?"
- Relational: "Is the heart size normal?"

The diversity of questions confirms that the models must learn both object detection (finding a mass) and attribute measurement (heart size), validating the difficulty of the task.

5.3 Feasibility and Expected Performance

Based on prior work on the VQA-RAD dataset and recent advances in biomedical vision-language models, we can establish realistic performance expectations for both approaches evaluated in this study.

The original VQA-RAD dataset paper reports that traditional CNN-based VQA architectures achieve relatively modest performance. Specifically, baseline models such as Stacked Attention Networks (SAN) (Yang et al., 2016) and Multimodal Compact Bilinear pooling (MCB) (Fukui et al., 2016) obtain accuracy scores in the range of approximately 55%-61% on closed-ended questions, highlighting the difficulty of radiology-focused visual question answering when using conventional feature-fusion pipelines (Lau et al., 2018). These results provide a reasonable lower-bound expectation for the ResNet + BiLSTM baseline implemented in this project, particularly given our stricter scope (Chest X-rays only) and image-level data splitting strategy.

In contrast, recent biomedical Vision-Language Models demonstrate substantially stronger performance. LLaVA-Med, a domain-adapted extension of the LLaVA framework fine-tuned on large-scale biomedical image-text data, reports closed-set accuracy exceeding 80% on the VQA-RAD benchmark after task-specific fine-tuning (Li et al., 2023). While zero-shot performance of general-purpose VLMs on medical data is often near random due to domain mismatch, the incorporation of biomedical pretraining and parameter-efficient fine-tuning techniques such as LoRA is expected to significantly improve results.

Therefore, we hypothesize that the fine-tuned LLaVA-Med model will achieve accuracy in the range of 75%-82% on the binary Chest X-ray VQA subset, substantially outperforming the CNN-based baseline. This expected performance gap directly motivates the comparative analysis, allowing us to quantify whether the increased architectural complexity of a Vision-Language Model yields meaningful gains for narrowly scoped, clinically relevant binary reasoning tasks.

5.4 Conclusion of Preliminary Phase

The preliminary analysis confirms that:

1. The subset is viable for training (balanced classes).
2. The leakage-proof splitting strategy is ready to be deployed.
3. The computational plan (using LoRA for the VLM) is feasible within the provided cloud resource constraints.

The next phase involves the active training of the ResNet baseline followed by the fine-tuning of LLaVA-Med.

References

Chiang, W.-L., Li, Z., Lin, T., Sheng, Y., Wu, H., Zhang, Z., … Stoica, I. (2023, March 30). Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality*. LMSYS Blog. (LMSYS)

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 457-468). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1044 (ACL Anthology)

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778). IEEE. https://doi.org/10.1109/CVPR.2016.90 (DBLP)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735 (PubMed)

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In ICLR 2022 (Poster). OpenReview. (OpenReview)

Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5, Article 180251. https://doi.org/10.1038/sdata.2018.251 (Nature)

Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). Visual Question Answering in Radiology (VQA-RAD) [Dataset]. Open Science Framework. https://doi.org/10.17605/OSF.IO/89KPS (OSF)

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In NeurIPS 2023: Datasets and Benchmarks Track.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. arXiv:2304.08485. (arXiv)

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162 (ACL Anthology)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on

Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 139, pp. 8748-8763). PMLR. (Proceedings of Machine Learning Research)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115, 211-252. https://doi.org/10.1007/s11263-015-0816-y (Springer)

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681. https://doi.org/10.1109/78.650093 (ACM Digital Library)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971. (arXiv)

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 21-29). IEEE. https://doi.org/10.1109/CVPR.2016.10 (Proceedings of Machine Learning Research)