Falcon Medical Visual Question Answering

Ameera Bawazir, Hassan Alshanqiti, Kebin Wu, Fatima Albreiki

Technology Innovation Institute
9639 Masdar City, Abu Dhabi, United Arab Emirates
Ameera.Bawazir@tii.ae, hassan.alshanqiti@tii.ae, kebin.wu@tii.ae, fatima.albreiki@tii.ae

Abstract

Vision-Language Models (VLMs) bridge the gap between visual and textual data, enabling multimodal tasks like Visual Question Answering (VQA). Leveraging this capability, Medical VQA systems have the potential to transform clinical decision-making by allowing healthcare providers to query medical images—such as X-rays, MRIs, and CT scans—and receive rapid, informed responses, thereby speeding up diagnoses and treatment planning. In this work, we introduce Falcon Med-VQA, a generative VQA system meticulously designed to interpret visual and textual medical data and generate free-form answers to medical questions. By leveraging a vision language model and a dynamic model selection mechanism, Falcon Med-VQA ensures relevance and precision in its responses. The system is equipped with an intuitive user interface that displays top answers with Confidence Scores (CF), enhances explainability through medical terminology extraction, and offers attention map visualizations for improved interpretability. Our experiments demonstrate that Falcon Med-VQA achieves comparable performance against specialized models and outperforms recent generative approaches in a key benchmark.

Introduction and Related Work

Vision-Language Models (VLMs) (Radford et al. 2021), (Jia et al. 2021), (Alayrac et al. 2024), (Li et al. 2022) bridge the gap between visual and textual modalities by combining vision and language components. These models, trained on extensive datasets of paired text and images, enable them to understand the relationship between visual content and textual information. This capability allows them to perform tasks such as Visual Question Answering (VQA), where they interpret images and generate relevant textual responses to natural language queries. However, despite significant advances, most VLMs are predominantly trained on natural images and language data (Li et al. 2021), (Liu et al. 2024b), which limits their effectiveness in the medical domain. To address these challenges, Medical Visual Question Answering (Med-VQA) models have been developed, specifically designed to interpret medical visual content and provide accurate responses to medical queries, tackling the unique demands of biomedical questions.

Traditionally, Medical VQA has been treated as a multi-class classification problem, where models select an answer from a predefined set based on joint image-question representations (Li et al. 2023a)(Li et al. 2023b) (Eslami, Meinel, and De Melo 2023). While effective, this approach is limited by the finite set of answers, severe class imbalances, and unseen vocabulary, making it less ideal for free-form medical questions. Recently, decoder-based vision-language models have gained attraction, enabling generative solutions that produce open-ended responses and better handle diverse and complex queries (Li et al. 2024) (Zhang et al. 2023) (Liu et al. 2024a).

To this end, we introduce Falcon Med-VQA, a system meticulously designed to excel in interpreting visual and textual medical data while generating insightful answers to visual medical questions. The system is named after the Falcon LLM (Malartic et al. 2024), which is used as the language decoding component in our system. Our contributions are summarized as follows:

- Generative Learning Framework: Falcon Med-VQA treats medical VQA as a generative task, using a vision-language model to produce free-form answers.
- Dynamic Model Selection: The system selects the optimal VQA model based on the medical image type, leveraging a medical image classifier to improve accuracy.
- Intuitive User Interface: It provides an intuitive user interface by presenting the top three answers with confidence scores, aiding clinical decision-making.
- Enhanced Explainability: Our system highlights key medical terms from the top answer and provides explanations for clear understandings of the context.
- Attention Map Visualization: Falcon Med-VQA generates attention map visualizations, highlighting the specific areas of the medical image the model focuses on, offering insights into its decision-making process.

To validate our approach, we conducted extensive experiments on publicly available medical VQA benchmark datasets. The results demonstrate that Falcon Med-VQA not only delivers performance comparable to highly specialized medical VQA models but also surpasses the latest generative models on key benchmarks.

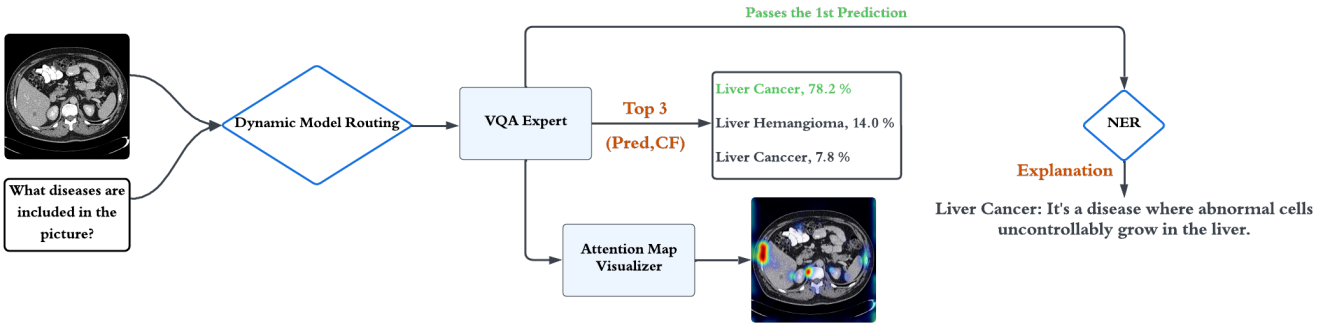

Figure 1: Overview of Falcon Med-VQA system. The process flow includes user input, dynamic model routing, VQA expert, top answer and confidence generation, attention map visualization, named entity recognition model, and explanation generation.

Falcon Med-VQA System Overview

Falcon Med-VQA Model

The Falcon Med-VQA model architecture integrates three key components to address medical visual question answering tasks. For visual feature extraction, we employ PMC-CLIP (Lin et al. 2023), an encoder optimized for medical image analysis. The connection component, which projects the visual representation into the language latent space, is implemented with a multi-layer transformer decoder with learnable query embeddings (Zhang et al. 2023). For language processing and response generation, we leverage the Falcon-11B model (Malartic et al. 2024). Our model architecture efficiently processes medical images, fuses multimodal information, and generates accurate, context-aware responses to medical queries. By combining these specialized components, Falcon Med-VQA achieves high performance across diverse medical VQA tasks.

To optimize performance in medical VQA tasks, the training consists of three stages. In the first two stages, we train the models with PMC-VQA dataset (version 2) [1], which contains approximately 152K image-caption and image-question pairs. In the **first stage**, the model undergoes pre-training on image-caption pairs to align visual-linguistic multimodal medical concepts. In the **second stage**,it is fine-tuned on medical VQA pairs. This stage refines the model's ability to understand and answer specific medical questions based on visual input, enhancing its performance on medical VQA tasks. In the **third stage**, we fine-tune the model on the SLAKE (Liu et al. 2021) dataset for radiology-specific tasks. For the pathology-specific model, we fine-tune on PathVQA (He et al. 2020), SLAKE, and VQA-RAD (Lau et al. 2018), covering both pathology and radiology images to build expertise in both fields.

During training, we utilize Low-Rank Adaptation (LoRA) (Hu et al. 2021) for the LLM component while performing full fine-tuning on the vision encoder and the connection component. This strategy enables efficient fine-tuning, allowing the model to adapt to domain-specific visual-linguistic tasks while preserving the LLM's pre-trained knowledge and reducing computational overhead.

Falcon Med-VQA performs comparably with MedVInT-TD (Zhang et al. 2023) and PefoMed on radiology and pathology benchmarks, significantly outperforming them on the SLAKE VQA task with 86.6% accuracy on open-set questions and 89.5% on closed-set questions.

Falcon Med-VQA User Interface

The workflow of Falcon Med-VQA system is shown in Figure 1. To start with, the user inputs a medical image and a related question. The image is then analyzed by a classifier powered by ClipMD (Glassberg and Hope 2023) to determine whether it belongs to pathology or radiology, followed by dynamic model routing to select specialized checkpoints optimized for each domain, ensuring optimal performance for these distinct medical imaging domains. Once the appropriate checkpoint is chosen, the system generates the top three possible answers, each accompanied by an associated CF. Additionally, an attention map is produced using the Grad-CAM (Selvaraju et al. 2017) technique, highlighting the critical areas of the medical image that influence the model's predictions. The system then focuses on the top-ranked answer and applies DeBERTa-Med-NER-2 (Mattupalli 2024), a Named Entity Recognition (NER) model, to identify specific medical terms within the response. These identified terms are subsequently passed to GPT-4 (Achiam et al. 2023), which provides detailed explanations, enabling an end-to-end solution even for users without specialized medical knowledge. On average, user requests are processed within 4.9 seconds in our interface, ensuring fast, efficient responses and prompt access to results.

Conclusion and Future Work

We introduce Falcon Med-VQA, a generative visual question answering system designed for medical applications. It demonstrates competitive performance in interpreting medical images and answering complex queries, while the user-friendly interface further enhances its clinical utility. Future work will focus on expanding its capabilities to include comprehensive medical report generation, assisting healthcare professionals in diagnosis and treatment planning.

[1]This version of the dataset can be found at https://huggingface.co/datasets/xmcmic/PMC-VQA

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millicah, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2024. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Eslami, S.; Meinel, C.; and De Melo, G. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, 1181–1193.

Glassberg, I.; and Hope, T. 2023. Increasing Textual Context Size Boosts Medical Image-Text Matching. *arXiv preprint arXiv:2303.13340*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S. R.; Xiong, C.; and Hoi, S. C. H. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Neural Information Processing Systems*.

Li, P.; Liu, G.; He, J.; Zhao, Z.; and Zhong, S. 2023a. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 374–383. Springer.

Li, P.; Liu, G.; Tan, L.; Liao, J.; and Zhong, S. 2023b. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 525–536. Springer.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Liu, G.; He, J.; Li, P.; He, G.; Chen, Z.; and Zhong, S. 2024a. PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging. arXiv:2401.02797.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Malartic, Q.; Chowdhury, N. R.; Cojocaru, R.; Farooq, M.; Campesan, G.; Djilali, Y. A. D.; Narayan, S.; Singh, A.; Velikanov, M.; Boussaha, B. E. A.; et al. 2024. Falcon2-11B Technical Report. *arXiv preprint arXiv:2407.14885*.

Mattupalli, S. 2024. Medical-NER. https://huggingface.co/blaze999/Medical-NER. Accessed: 2024-09-13.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv preprint arXiv:2305.10415*.