Proof-of-Concept: *From Monolith to Manifold: Collaborative Medical Visual Question Answering*

Xuelong An
ucabxan@ucl.ac.uk

Abstract

This is a proof-of-concept project where we propose a pipeline for medical visual question answering on the MEDIQA-M3G benchmark on dermatology cases. It is inspired by literature on multi-agent large language models which leverages multiple, interoperable models which can discuss with each other. Our pipeline is highly customizable, and can be tailored for various purposes, such as adopting parameter-efficient finetuning techniques in low-resource settings, or performing retrieval augmented generation for knowledge-based answer generation.

1 Introduction

1.1 Motivation

Dr. Eric Topol once remarked that artificial intelligence models are not meant to replace doctors, but rather assist them in their tedious workloads so that they can rekindle their humane connection with patients (Topol, 2019). Indeed, the advent of deep learning (DL) continues to provide unprecedented opportunities for clinicians to streamline their workflows in the many aspects of public healthcare.

One such area is telehealth, defined as remote patient care delivery through electronic means such as online visits. It provides opportunities such as cost-effective communication means between patient and clinician, but also brings new challenges such as increased workloads for physicians who also have to balance office visits (Bishop et al., 2013), or ensuring the quality of electronic visits can match their in-person counterparts. Such challenges resonates with us until this day, especially given global COVID 19 epidemic which accelerated the maturation of digital health portals (Yim et al., 2024).

The adoption of Artificial Intelligence (AI) technologies such as DL helps address many of the challenges surrounding telehealth. In particular for tasks like automatic response generation, it has prospects of accurately and timely giving patients a first-hand diagnostic account of their condition, and/or giving physicians suggestions to their patients workflows, thus alleviating work burden (Yim et al., 2024). Furthermore, AI adoption can address the shortage of medical personnel, especially given the World Health Organization (WHO) estimated that over 45% of the countries across the globe have less than one physician available per 1000 population (Organization, 2019).

There is much prior work in DL for automatic response generation in healthcare, such as in multimodal visual question answering (VQA), which aims at response generation through a combination of imaging and text modalities. The MEDIQA-2024 Multilingual & Multimodal Medical Answer Generation (M3G) Shared Task is a benchmark aimed at evaluating multimodal approaches. It focuses on visual question-answer generation localized to dermatology cases, evaluated in 3 languages: English, Chinese, and Spanish (Yim et al., 2024).

There are many approaches to solve the benchmark, such as prompting proprietary multimodal models through API calls, or engineering a pipeline and training them on the MEDIQA-2024 dataset. In the former approach, a plethora of powerful vision-language models have been proposed over the years, which can be classified as either 1) API models like the families of GPT-4, Claude and Gemini, or 2) open-sourced models like LLAVA, Pixtral, Qwen, Molmo, among others (Deitke et al., 2024). The latter approach, albeit more difficult, enables more flexible exploration of model architectures, and addressing practical clinical concerns such as memory-efficiency and deployment in edge-devices. In this work, we propose a pipeline as a proof-of-concept for medical visual question answering.

1.2 Literature Review

Open-ended answer generation: Prior work in medical-VQA often focus in structured answers given a question, such as choosing the correct answer from multiple choices. Because the answers follow a strict syntax, this essentially treats VQA as a text classification task. The alternative consists of free-form answer generation, with approaches such as MedFuseNet which consists of a general pipeline of image feature extraction, question feature extraction, a feature fusion module, and an answer prediction module (Sharma et al., 2021). These general blueprint underlies much of the multimodal VQA architectures, and one can experiment with different existing models to fit in the pipeline.

Parameter-efficient finetuning (PEFT): Training each of the above blueprint components from scratch can be prohibitively expensive and impractical given the plethora of readily available open-source models which have been pretrained on a large corpora of biomedical data. There exists many resource-efficient finetuning methods, like *last-layer transfer*, where earlier layers of a model (often encoders, or transformers) remain frozen, while only last layers are learnable. One can also add learnable matrices to an existing architecture or to the learning method, and only fine-tune these additional matrices; differences in how these matrices are added to the architectural components and where to add them in the optimization procedure lead to different finetuning approaches like LoRA, prefix tuning (van Sonsbeek et al., 2023), sparse adapter (Hao et al., 2024), or prompt tuning of the language model component (Lester et al., 2021).

Society of minds: The general trend in DL research is to propose a single architecture which can achieve state of the art (SoTA) performance on a well established benchmark. Different proposed models carry their advantages and limitations. For example, a VQA model with a vision encoder pretrained on rare skin lesions may answer queries more accurately from patients with these edge cases, but perform badly on common lesions. Another VQA model (which could be an ablation of the former) may perform better at spotting normal skin lesions and perform badly on rare diseases. A question which arises is how can we combine their strengths and address each other's limitations, akin to members of a team compensating for each other's level of expertise. One approach is to combine multiple VQA models to leverage their interoperability so that they can discuss with each other (Zhuge et al., 2023), transitioning from a monolith model to a manifold of models. Since medicine is inherently a highly collaborative field, we can draw inspiration from the literature on multi-agent large-language models (LLMs) to assemble various VQA models to cooperate with one another to answer a patient's query, instead of resorting to the "SoTA" candidate. Research on multi-agent LLMs has shown various benefits of building this collaborative environment, such as improving factuality, and debating to ensure consistency of ideas (Guo et al., 2024).

2 Methodology

The MEDIQA-M3G is a benchmark collected from Chinese dermatology cases sourced from AiAiYi. For each anonymized patient, we're provided a stack of dermatological images $\mathbf{I}$, a unique input question in natural language $\mathbf{Q}$, along with a list of answers $\mathbf{A}$ given by queried specialists. Both question and answers are provided in Chinese, English and Spanish, where the latter languages are GPT-4 based translations in the train-split, and medical expert translations in the test-split. For some answers, we're provided rankings based on the expertise of the specialists, reflecting their credibility. There is a variable amount of images and answers within and across patients. From a technical perspective, this constrain us to use a batch size of 1 as datapoints can not be stacked. Furthermore, for training purposes, we apply a preprocessing step where ensure there is the same amount of answers and images per patient. If a case has more images than answers, we upsample answers by copying the longest answers assuming they contain more information such as a justification for a diagnosis, and thus is of higher quality. If a case has more answers than images, then we augment images by applying random data augmentations techniques such as horizontal flip, changing brightness, contrast, rotation, resizing, cropping, blurring, or adding random Gaussian noise (Figure 1).

Given this context, we propose the following proof-of-concept framework for addressing medical VQA. Figure 2 depicts our training regime. For each VQA model of the society, we aim to find optimal parameters θ^* for a model by maximizing the conditional log-likelihood $\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(\mathbf{A}_i|\mathbf{Q}, \mathbf{I})$. As an example,

(Augmented) Images:

(Augmented) Answers:

English: ['Eczema', 'The possibility of psoriasis is high.', 'Eczema with infection', 'The likelihood of pustular psoriasis is high.']
Chinese: ['蒙 湿疹', '银屑病的可能性大', '湿疹并感染', '脓疱型银屑病可能性大']
Spanish: ['Eczema', 'La posibilidad de psoriasis es alta.', 'Eczema con infección', 'La probabilidad de psoriasis pustulosa es alta.']

Question:

English: Young male, 10 years old. The condition has been distributed symmetrically since onset. It improved with medication, but recently it has been recurring. Is it psoriasis? If so, how should it be treated further? Thank you!!
Chinese: 少年男性，10岁。发病以来成对称分布。曾用药好转。但最近又反复。是不 是银屑病？ ？ 如果是进一步如何治疗？ ？ ？ 谢谢！ ！
Spanish: Varón joven, 10 años de edad. La condición ha sido distribuida simétricamente desde su inicio. Mejoró con la medicación, pero recientemente ha estado reapareciendo.¿Es psoriasis? Si es así, ¿cómo debería tratarse más adelante? ¡Gracias!

Figure 1: Depiction of a patient case in the MEDQA-M3G Shared Task. In this example, there were more answers than images. Two of the screenshots are augmented from a random sample of images, and this is done to match the amount of images with answers per batch for technical purposes.

we use the language models employed by (van Sonsbeek et al., 2023): GPT-2XL (red), BioGPT (green) and BioMedLM (blue) as question feature encoders and answer prediction modules. We prompt them such that each is assigned a medical role (e.g. resident or dermatologist) and learn to answer the patient query from different perspectives. For image feature encoders we can leverage CLIP, although for dermatology we can also use more localized feature embedders like Derm Foundation (Steiner and Pilgrim, 2024; Rajeev V et al., 2024), which has higher prospects of outputting features for generalization across patients. The architecture follows the general blueprint mentioned in (Sharma et al., 2021). Accounting for resource-constrained settings, many PEFT methods can be applied in addition to model compression, like quantization of the language models, along with mixed-precision training, where we decrease the bitwise precision of the model parameters from an original 64-bits to 16-bits.

After training multiple VQA models, during inference, we follow the pipeline shown in Figure 3. Each VQA model receives a test query and offers a list of answers, which may differ between one another given they may have learnt differently and focused on different aspects of the image embeddings during training. We provide this list of answers and the original question to a LLM such as Mistral-7b and ask for a concise answer given the different diagnoses. For evaluation, multiple metrics which account for answer quality, multiple languages, among other aspects are employed. These include BERTScore, deltaBLEU (a variant of SacreBLEU) and MEDCON (Yim et al., 2024).

3 Discussion

Our training and inference regime is similar to (García and Lithgow-Serrano, 2024), where the difference lies in that we use multiple VQA model responses, instead of responses from a single VQA model, which are summarized by the LLM.

For the inference pipeline, we can also leverage retrieval augmented generation (RAG) from dermatology ontologies to improve the medical accuracy of answers.

For proprietary models which can't be finetuned, we can instantiate the inference pipeline by replacing placeholders for language models with API calls with LLMs like SoTA, multimodal LLMs like Claude-3 or GPT-4o to explore the above roleplaying, collaborative prompting strategy. The drawback of such approach is that querying multiple proprietary models incur additional costs, as we are unable to apply model compression techniques if our goal is deployment over edge devices.

As the field of multimodal deep learning progresses, we also observe a trend of open-source software, thus models like Molmo can be leveraged and used in the proposed training and inference framework above.

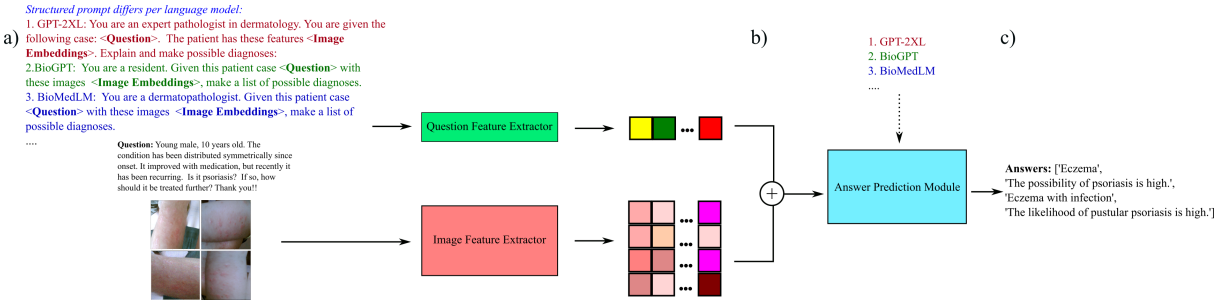

Figure 2: Illustration of our customizable training pipeline. At a) we extract features from the question **Q** which is constructed differently per each language model, and from the stack of images **I**. At b) we fuse the features to be passed to the language model for free-form answer prediction. At c), each VQA model outputs an answer via greedy search (or beam search with a width of 1), and is trained by minimizing the cross-entropy loss function between ground-truth tokens and the predicted tokens. In the illustration we depict prompts in Spanish.

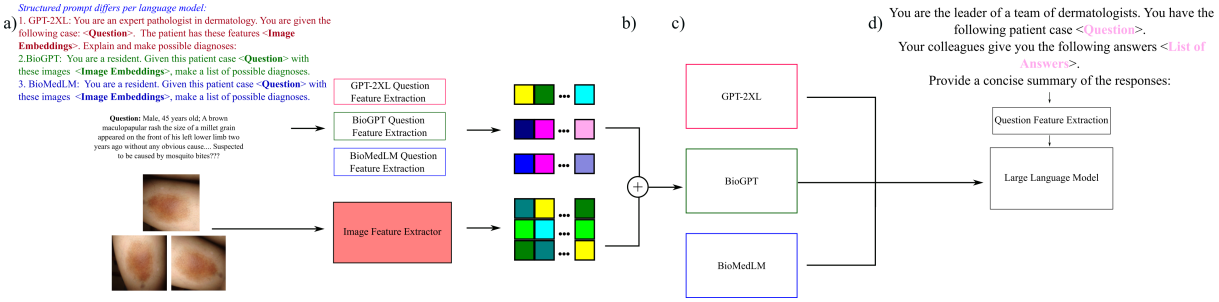

Figure 3: Illustration of our customizable inference pipeline. Similar to training's stage a), we extract features from the question and images, and fuse them at b) so that at c) each of the language models output a list of answers. At d), the answers of each VQA model is put as context for a prompt to a LLM, where we query for a concise answer that accounts for the different diagnoses offered by each model. In the illustration we depict prompts in Spanish.

References

Tara F. Bishop, Matthew J. Press, Jayme L. Mendelsohn, and Lawrence P. Casalino. 2013. Electronic communication improves access, but barriers to its widespread adoption remain. *Health Affairs*, 32:1361–1367.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models.

Ricardo García and Oscar Lithgow-Serrano. 2024. Neui at mediqa-m3g 2024: Medical vqa through consensus. *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges.

Jitai Hao, Weiwei Sun, Xin Xin, Qi Meng, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Meft: Memory-efficient fine-tuning through sparse adapter. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2388.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

World Health Organization. 2019. Global health workforce statistics database.

Rikhye Rajeev V, Aaron Loh, Grace Eunhae Hong, Preeti Singh, Margaret Ann Smith, Vijaytha Muralidharan, Doris Wong, Rory Sayres, Michelle Phung, Nicolas Betancourt, Bradley Fong, Rachna Sahasrabudhe, Khoban Nasim, Alec Eschholz, Basil Mustafa, Jan Freyberg, Terry Spitz, Yossi Matias, Greg S Corrado, Katherine Chou, Dale R Webster,

Peggy Bui, Yuan Liu, Yun Liu, Justin Ko, and Steven Lin. 2024. Closing the ai generalization gap by adjusting for dermatology condition distribution differences across clinical settings.

Dhruv Sharma, Sanjay Purushotham, and Chandan K. Reddy. 2021. Medfusenet: an attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11.

Dave Steiner and Rory Pilgrim. 2024. Health-specific embedding tools for dermatology and pathology.

Eric J. Topol. 2019. *Deep medicine : how artificial intelligence can make healthcare human again*. Basic Books.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 726–736.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024. Overview of the mediqa-m3g 2024 shared task on multilingual multimodal medical answer generation. *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Al Kader, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. *Workshop on robustness of zero/few-shot learning in foundation models*.