LTRC-IIITH at MEDIQA-M3G 2024: Medical Visual Question Answering with Vision-Language Models

Jerrin John Thomas, Sushvin Marimuthu, Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad, India
{jerrin.thomas, sushvin.marimuthu}@research.iiit.ac.in
{param.krishna}@iiit.ac.in

Abstract

In this paper, we present our work to the MEDIQA-M3G 2024 shared task, which tackles multilingual and multimodal medical answer generation. Our system consists of a lightweight Vision-and-Language Transformer (ViLT) model which is fine-tuned for the clinical dermatology visual question-answering task. In the official leaderboard for the task, our system ranks 6th. After the challenge, we experiment with training the ViLT model on more data. We also explore the capabilities of large Vision-Language Models (VLMs) such as Gemini and LLaVA.

1 Introduction

The rapid evolution of telecommunication technologies, coupled with increased healthcare demands and the recent challenges posed by the pandemic, has accelerated the adoption of remote clinical diagnosis and treatment. Alongside conventional live consultations conducted via telephone or video, asynchronous methods such as e-visits, emails, and messaging chats have emerged as practical and cost-effective alternatives.

This task (wai Yim et al., 2024a) focuses on addressing the challenge of generating suitable textual responses to queries in clinical dermatology, taking into account multimodal inputs such as clinical history, queries, and accompanying images.

This paper describes our proposed solution. We fine-tune ViLT (Kim et al., 2021) for the visual question-answering task with the training data provided for the challenge. We choose ViLT due to its lightweight nature and ability to handle both visual and textual inputs efficiently. After the challenge, we also explore how large Vision-Language Models (VLMs) such as Gemini (Team, 2024) and LLaVA (Liu et al., 2023) perform in this task. These models stand at the top of the multi-modal benchmarks such as Massive Multi-discipline Multimodal Understanding and Reasoning benchmark (MMMU) (Yue et al., 2023).

2 Related Work

Previous research has predominantly focused on consumer health question-answering but has been limited to textual inputs (Ben Abacha et al., 2019). Similarly, existing work on visual question-answering has primarily concentrated on radiology images, lacking integration with additional clinical text inputs (Abacha et al., 2019). Moreover, while significant research has been conducted on dermatology image classification, the emphasis has largely been on lesion malignancy classification for dermatoscopy images (Li et al., 2022).

Recently, there has been a surge in the development of multimodal models, particularly large vision-language models (VLMs). These models integrate both textual and visual information, allowing them to understand and generate content that combines both modalities. VLMs typically employ techniques such as joint embedding to unify the representations of text and images in the same embedding space. During training, they utilize datasets that contain interleaved text and images, enabling the model to associate textual descriptions with visual content effectively. This process enables VLMs to grasp nuanced relationships between words and visual elements, facilitating tasks like visual question-answering (VQA). In VQA, these models can accurately respond to questions about images by understanding the content of both the image and accompanying text, showcasing their ability to comprehend and synthesize information across modalities.

For the medical domain, VLMs have been trained on medical corpora and developed for various clinical tasks. MedBLIP (Chen et al., 2023) aids computer-aided diagnosis (CAD) in the medical field. It tackles the challenge of combin-

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 703–707
June 21, 2024 ©2024 Association for Computational Linguistics

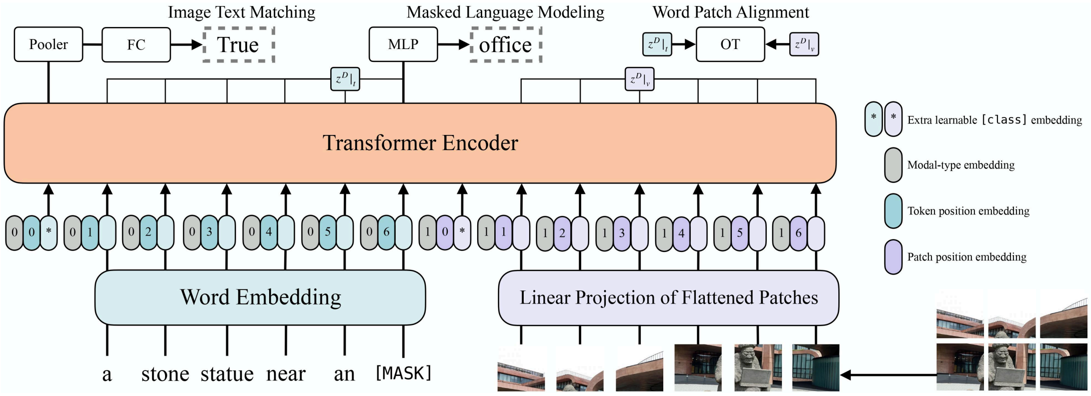

Figure 1: ViLT Model Architecture (from Kim et al. (2021))

ing image and text data from electronic health records for medical diagnosis. The model shows promising results in classifying healthy, mildly impaired, and Alzheimer's patients and also demonstrates the ability to answer medical questions based on visual information. PMC-LLaMA (Wu et al., 2024), designed specifically for medical applications, demonstrates superior performance on medical question-answering tasks. Med-Flamingo (Moor et al., 2023) is a model that can learn from small datasets by embracing in-context learning for the multi-modal medical domain. BiomedGPT (Zhang et al., 2024) is a unified model designed to handle diverse medical data and perform various tasks such as diagnosis and summarization. LLaVA-Med (Li et al., 2023) utilizes a massive dataset of biomedical images and captions from PubMed Central and employs the powerful language model GPT-4 to create diverse training examples.

3 Dataset

The dataset provided for the shared task is DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology (wai Yim et al., 2024b). It is translated and adapted from Chinese telemedicine datasets. The given dataset consists of 842 samples in the training set, 56 for validation and 100 for testing. Each sample has a query, clinical history, and one or more associated images. The textual content is provided in three languages - Chinese, English, and Spanish. The English and Spanish versions of the training set are generated through machine translation from the Chinese original. The validation and test datasets are manually translated by human translators. This approach allows for comprehensive testing and validation of models across multiple languages while ensuring the integrity and quality of the data through both machine and human translations.

Dataset	Size
Train	842
Valid	56
Test	100

Table 1: Dataset Splits

4 System Description

Our system is made of a fine-tuned Vision-and-Language Transformer (ViLT) (Kim et al., 2021) model. ViLT is lightweight and can handle data of both textual and visual modalities.

4.1 Model Description

ViLT is a pre-trained multimodal model that simplifies the processing of visual inputs by treating them in the same convolution-free manner as text inputs. This approach reduces the computational complexity of the model-specific components compared to the transformer component for multimodal interactions. ViLT is pre-trained on three objectives: image-text matching, masked language modeling, and word-patch alignment. For the visual question-answering task, we employ a ViLT model with a classifier head on top, which consists of a linear layer applied to the final hidden state of the [CLS] token. This architecture allows the model to leverage its pre-trained multimodal representations to effectively answer questions about visual inputs. The key advantages of ViLT are its simplicity and computational efficiency

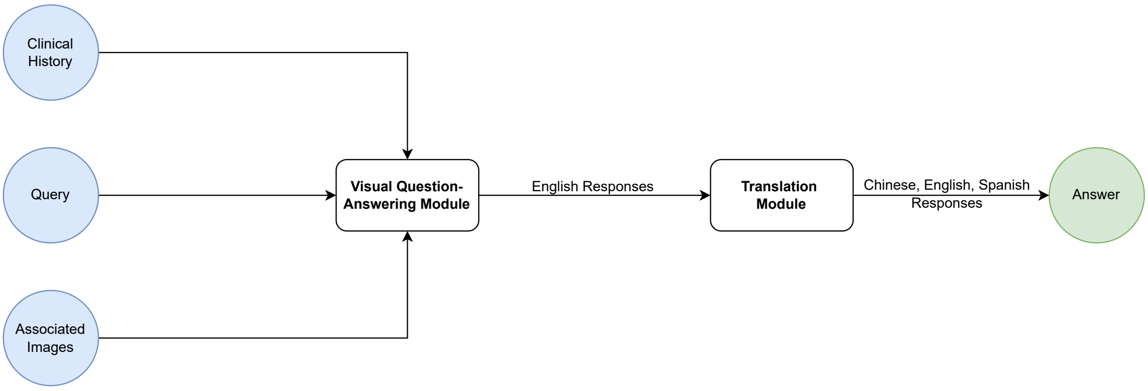

Figure 2: After-Challenge System Workflow

4.2 Data Pre-Processing

In the pre-processing step, we focus solely on the English content, ignoring the Chinese and Spanish content. We select specific fields from the data, namely `image_ids`, `query_title_en` for questions and `query_content_en` for labels. We proceed to structure the dataset by flattening it and organizing it into tuples containing the image IDs, questions, and labels.

Following the dataset flattening, we encode both the images and texts using the ViltProcessor, a processor tailored for our model. This encoding step is crucial for transforming the raw textual and visual inputs into formats suitable for ingestion by the model. By leveraging the capabilities of the ViltProcessor, we ensure that the data is prepared as required for the subsequent training process. With the pre-processing complete, we obtain a refined and standardized dataset ready for training our model on visual question-answering tasks.

Initially, we build a dataset with only 200 samples and after the challenge, we use all 842 samples for training. We process the data in batches of 200 samples each and merge all the processed data at the end. This approach allows us to effectively manage memory usage without sacrificing the richness of our dataset, ensuring robust model training and analysis.

4.3 Fine-Tuning

Leveraging the ViLT Processor, we seamlessly load our data into the model and the ViLT model is fine-tuned using the processed dataset. During fine-tuning, we tune the hyperparameters to suit our objectives effectively. The batch size is set to 4 and the learning rate at $5e-5$. We train the model for 10 epochs. This fine-tuning process allows the ViLT model to adapt and specialize to the nuances of the specific visual question-answering task, ensuring that it can effectively comprehend and respond to questions of the clinical dermatology domain. With these hyperparameters in place, we aim to achieve optimal performance and robustness in our model's ability to answer questions accurately and comprehensively.

4.4 Inference

Following the completion of the fine-tuning, we perform inference on the model. We follow the same pre-processing steps. We utilize `encounter_id` as the question identifier, `query_title_en` as the question itself, and `image_ids` as additional contextual information. Leveraging the model's learned representations and understanding of visual and textual inputs, we generate predictions for each sample.

Once the predictions are obtained, we format the results into a JSON file, organized as an array of JSON objects. Each JSON object contain `encounter_id` as a unique identifier and a responses array. Within this array, we include predicted responses in English (`content_en`), leaving the corresponding fields for Chinese (`content_zh`) and Spanish (`content_es`) empty, as our training and prediction efforts were focused solely on the English language.

5 After-Challenge Experiments

After the challenge, we develop a system containing two modules - visual question-answering module and translation module. We experiment with two models - Gemini 1.0 Pro Vision (Team, 2024)

Models	Chinese		English		Spanish	
	DeltaBleu	BERTScore	DeltaBleu	BERTScore	DeltaBleu	BERTScore
ViLT (200 Samples)	-	-	0.46	0.83	-	-
ViLT (842 Samples)	-	-	0.52	0.82	-	-
LLaVA - 1.6 34B	1.60	0.64	0.53	0.82	0.88	0.76
Gemini 1.0 Pro Vision	2.70	0.59	0.86	0.70	1.39	0.66

Table 2: Evaluation Results for Different Experiments

and Llava-1.6 34B (Liu et al., 2024).

5.1 Model Description

Gemini 1.0 Pro Vision is capable of comprehending inputs from both textual and visual sources, which can be both images or videos, yielding contextually relevant textual outputs. Serving as a foundational model, It excels across a spectrum of multimodal tasks, including visual comprehension, classification, summarization, and content generation from diverse visual inputs such as photographs, documents, infographics, and screenshots.

LLaVA-1.6 34B is a large vision-language model that stands out for its ability to understand and process both text and visual data, making it highly capable in general-purpose visual and language tasks. It is an auto-regressive language model, based on the transformer architecture, and has 34 billion parameters. It is fine-tuned on multi-modal instruction following data. LLaVA-1.6 34B has even surpassed the performance of models like Gemini Pro on some benchmarks.

5.2 Visual Question-Answering Module

> **VQA Prompt**
>
> You are a clinical dermatology assistant who can generate clinical responses, given the clinical history and a query, along with one or more associated images. Be concise and do not give additional information other than answering the query.
> {Associated Images}
> {Clinical History}
>
> {Query}

For the visual question-answering (VQA) module, we process the dataset and extract only the `encounter_id`, English content, and `image_ids` from the responses. Subsequently, we employ the model to predict results using a prompt created from the clinical history, query, and associated images (the prompt template is provided). These predictions are then stored in a JSON format, associating each encounter_id with a responses array containing the predicted English data. As for Chinese and Spanish content, we leave those fields empty, reflecting our current focus on English language prediction.

5.3 Translation Module

For the translation module, we use Gemini 1.0 Pro. The prompt template only has a simple instruction - "Translate from English to {language}". The English responses are translated into both Spanish and Chinese languages.

6 Evaluation Metrics

The evaluation process utilizes deltaBLEU (Galley et al., 2015), a metric that accounts for multiple correct responses. These responses are weighted based on various factors, including completeness, consistency with the most commonly provided answer as determined by human assessment, as well as author rank level and author validation level. The completeness metric assigns a score on a scale of {0.0, 0.5, 1.0}, indicating the extent to which the original query's question was addressed. A score of 1.0 signifies a fully answered query, 0.5 indicates a partial response, and 0.0 reflects a lack of response. If the query doesn't explicitly specify, it's assumed to seek information on both the disease and its treatment. Contains Most Frequent Answer rating is given on a scale of {0.0, 1.0}. A score of 1.0 is assigned if the response aligns with the most frequently provided answer.

7 Results

The evaluation results of the different experiments can be seen in Table-2. We submitted the system of ViLT trained on 200 samples for the challenge. Our submission ranks 8th in the leaderboard.

8 Conclusion

This work investigates vision-language models like ViLT, Gemini, and LLaVA for the challenging multilingual and multi-modal medical answer generation task in dermatology. A modular system with separate visual QA and translation components shows improved performance over the initial ViLT approach. A key strength is leveraging powerful multi-modal models that can effectively integrate visual and textual clinical data. Future efforts should focus on utilizing larger, more diverse datasets, incorporating stronger multi-modal reasoning, and rigorous evaluations by medical experts to ensure clinical utility and safety before real-world deployment. Overall, this mult-imodal approach holds promise but requires further advancements to be reliable for remote diagnosis and treatment.

References

Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CEUR Workshop Proceedings*, pages 9–12.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. 2023. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *Preprint*, arXiv:2305.10799.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Zhouxiao Li, Konstantin Christoph Koban, Thilo Ludwig Schenck, Riccardo Enzo Giunta, Qingfeng Li, and Yangbai Sun. 2022. Artificial intelligence in dermatology image analysis: Current developments and future trends. *Journal of Clinical Medicine*, 11(22).

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: A multimodal medical few-shot learner. ArXiv:2307.15189.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Wen wai Yim, Velvin Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024b. Dermavqa: A multilingual visual question answering dataset for dermatology. *CoRR*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, Hui Ren, Sunyang Fu, James Zou, Wei Liu, Jing Huang, Chen Chen, Yuyin Zhou, Tianming Liu, Xun Chen, Yong Chen, Quanzheng Li, Hongfang Liu, and Lichao Sun. 2024. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *Preprint*, arXiv:2305.17100.