A MULTIMODAL APPROACH TO DERMATOLOGY VQA

Sarthak Singh
Department of Natural Language Processing (IMS)
University of Stuttgart
Stuttgart, DE
st189880@stud.uni-stuttgart.de

Yassir El Attar
Department of Natural Language Processing (IMS)
University of Stuttgart
Stuttgart, DE
st191841@stud.uni-stuttgart.de

Keerthi Vasudevan
Department of Computer Science
University of Stuttgart
Stuttgart, DE
st189290@stud.uni-stuttgart.de

ABSTRACT

Artificial Intelligence has been making significant progress across various fields, not limited only to communication and text generation tasks, but also in healthcare; particularly in areas such as dermatology. Solutions such as image-based analysis have helped improve diagnostic accuracy. However, the integration of interactive systems capable of visual question answering (VQA) remains largely unexplored. This project aspires to extend the functionality of PanDerm, a dermatology-focused foundation model, by fine-tuning a language model on top of it resulting in a multi-model able to address VQA tasks. In the absence of dedicated VQA datasets for dermatology, we propose augmenting the ISIC 2018 dataset to create a VQA-compatible dataset. This augmentation enables the generation of tailored data, allowing PanDerm to learn associations between images, their diagnoses, and corresponding question-answer pairs. The project aims to detect skin conditions and there by provide assistance to dermatologists. We publish our code and details results with examples of generated answers on the github repository: https://github.com/asarthaks/FoundationModelsProject/

1 INTRODUCTION

This project aims to fine-tune large language models after fusing them with the PanDerm foundation model (Yan et al. (2024)) to perform VQA tasks. To address the lack of dedicated VQA datasets in dermatology, we integrated two complementary datasets: the ISIC 2018(Codella et al. (2018) and Tschandl et al. (2018)) dataset, which contains labeled skin images, and an information-seeking question-answer dataset based on dermatology. We constructed a training dataset by aligning the diagnostic labels from ISIC images with relevant questions and answers about similar diagnoses.
This project aims to improve PanDerm by enabling it to not only detect skin conditions but also provide answers to questions about symptoms, treatment options, and possible causes. The resulting system aspires to assist clinicians in analysing and diagnosing skin conditions. This paper outlines our methodology for dataset alignment, fine-tuning, and evaluation, highlighting the potential of foundation models to transform interactive healthcare solutions.

2 LITERATURE REVIEW

2.1 VISUAL QUESTION ANSWERING

Visual Question Answering (VQA) is a prominent area of research that focuses on enabling models to interpret visual inputs alongside textual queries to generate contextually accurate answers. In a typical VQA task, the model is presented with an image and a question and is expected to generate an answer by understanding the visual context of the image.

VQA tasks are commonly approached in two formats: (1) multiple-choice questions, where the model selects the correct option from predefined answers, and (2) open-ended questions, which require descriptive and unconstrained responses.

In recent years, several benchmarks have been developed to advance VQA research. For instance, Visual Dialog (Das et al. (2017)) extends traditional VQA by incorporating multi-turn dialog interactions. Similarly, datasets like TextVQA (Singh et al. (2019)) and ST-VQA (Biten et al. (2019)) focus on extracting textual information embedded within scenes in images.

Despite substantial progress in general VQA tasks, there has been limited exploration of VQA in medical domains like dermatology. Recently, the DermaVQA dataset (Yim et al. (2024)) was introduced as a benchmark for dermatology-specific VQA tasks. Their baseline model combines two components: an image-to-text generator (LLaVA-Med), which diagnoses skin conditions, and a text-to-text module powered by GPT-4 that generates user-friendly responses based on the diagnosis and query. Additionally, efforts such as IKIM at MEDIQA-M3G 2024 (Bauer et al. (2024)) demonstrated the potential to combine open source medical vision language models with large language models for cross-lingual visual question answering.

Building upon these advancements, we leverage PanDerm, a dermatology-focused foundation model (Yan et al. (2024)), to develop a visual-language model tailored for dermatology. Our approach incorporates lightweight language models and employs fusion techniques such as Cross-Attention, UNITER and ViLT to integrate visual and textual modalities effectively.

2.2 PANDERM

PanDerm is a multimodal foundation model designed for dermatology, leveraging self-supervised learning to process over 2 million skin disease images across four imaging modalities. Its architecture leverages techniques like the drop path regularization method, which enhances model robustness by stochastically dropping connections during training to prevent overfitting and improve generalization (Huang et al. (2016)). Its also integrates visual features with clinical context, enabling it to handle diverse dermatological tasks, including diagnosis, segmentation, and prognosis. PanDerm's ability to operate with minimal labeled data while also delivering state-of-the-art results makes it ideal for developing a VQA system tailored for dermatology, addressing the complexity of multimodal inputs and clinical queries (Yan et al. (2024)). Furthermore, we also looked at Gan et al. (2022) to better understand how to fuse vision language models and their architectures, particularly exploring dual encoders, fusion encoders, and cross-modal attention mechanisms to enhance multimodal integration.

3 DATASET

Given the limited availability of Visual Question Answering (VQA) datasets specific to dermatology, we created a tailored dataset to fine-tune PanDerm for VQA tasks. Our work primarily focuses on eight dermatological diagnoses: *Melanoma (MEL)*, *Vascular lesion (VASC)*, *Melanocytic nevus (NV)*, *Actinic keratosis (AK)*, *Benign keratosis (BKL)*, *Basal cell carcinoma (BCC)*, *Dermatofibroma (DF)*, and *Squamous cell carcinoma (SCC)*.

We primarily utilized the ISIC 2018 (Codella et al. (2018) and Tschandl et al. (2018)) dataset, which was also employed by PanDerm. Each image in the ISIC dataset is associated with a corresponding diagnosis. To create a VQA-compatible dataset, we augmented this dataset by generating generic questions corresponding to the diagnosis of each image, such as "What is the diagnosis in the image?" and "What are its causes?". Table 1 shows the Q&A data before augmentation, while Table 2 shows the data after the Q&A was used to augment the ISIC 2018 dataset.

The ISIC dataset originally contained 24,721 images of various dermatological conditions. To adapt it for our VQA task, we incorporated 101 unique question-answer pairs sourced from existing datasets and reliable online sources. By associating each image with multiple relevant questions, we expanded the dataset to 404,619 multimodal samples. Table 3 summarizes the distribution of images, Q&A pairs, and the final augmented dataset.

To construct this dataset, we employed a combination of techniques, sourcing data both from pre-existing datasets and web-based resources.

Table 1: Q&A data Before Augmentation

Prompt	**Response**
What is the etiology of melanoma?	This skin condition is Melanoma, caused by mutations in melanocytes producing melanin. UV exposure and tanning beds are major risk factors.
What is a Dermatofibroma?	A dermatofibroma is a harmless, nodular lesion under the skin, ranging from pink to brown in color.

Table 2: Expanded Dataset After Augmentation

Image	**Diagnosis**	**Prompt**	**Response**
	MEL	What is the etiology of this skin condition?	This skin condition is Melanoma, caused by mutations in melanocytes producing melanin. UV exposure and tanning beds are major risk factors.
	DF	What is the diagnosis in the image?	This is dermatofibroma, a harmless nodular lesion under the skin, ranging from pink to brown in color.

3.1 Augmenting with Existing Datasets

For *Melanoma (MEL)* and *Basal Cell Carcinoma (BCC)*, we leveraged a dermatology-based question-answer dataset available on Hugging Face (Mreeb (2024)). Using an iterative process for each diagnosis, we parsed through the dataset to identify prompts that directly referenced "melanoma" and "basal cell carcinoma," as well as their respective abbreviations (*MEL* and *BCC*). The selected prompts, along with their corresponding answers, were formatted into a JSON structure and integrated into our dataset file.

3.2 Augmenting with Web-Sourced Data

For the remaining diagnoses (*NV*, *AK*, *BKL*, *DF*, *VASC*, and *SCC*), we manually curated question-answer pairs from various web sources that are considered as a reliable dermatology-related inquiries sources (Anna Hernández (2024) and Altruderm (2024)). Frequently asked questions relevant to each diagnosis were collected and transformed into a VQA-compatible format. These included typical user queries regarding symptoms, treatment options, and diagnostics. The curated data was formatted into the same JSON structure as the entries in the section 3.1 and appended to the dataset.

4 Methodology

This project implements a multimodal Visual Question Answering (VQA) system by leveraging a cross-modal attention mechanism to combine visual and textual modalities. The system is designed to process input images along with textual queries to generate accurate textual answers.

4.1 Architecture

The model uses a Cross-Modal Attention mechanism built using PyTorch's MultiHeadAttention module (as showin in Figure 1). This method helps the system connect textual questions with the relevant parts of an image, enabling it to reason across both text and visuals. Specifically, the query originates from the textual input (question), while the keys and values are derived from the image.

The architecture is composed of the following components:

Diagnosis	Images (Original)	Q&A Pairs	Augmented Dataset (Images × Q&A)
MEL	4,522	32	144,704
NV	12,875	15	193,125
BCC	3,323	9	29,907
AKIEC	867	11	9,537
BKL	2,624	8	20,992
DF	239	16	3,824
VASC	253	10	2,530
Total	**24,721**	**101**	**404,619**

Table 3: Dataset statistics before and after augmentation

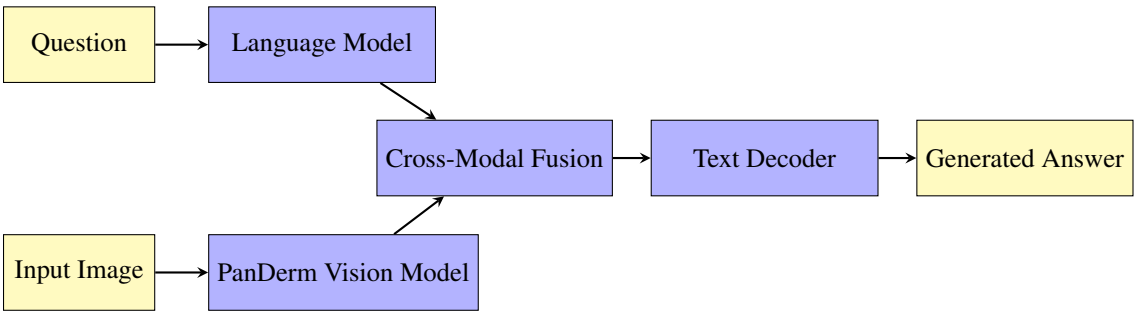

Figure 1: Architecture of the Visual Question Answering (VQA) Model. The model incorporates a visual encoder (PanDerm), a language model (BERT), and a decoder (GPT) with cross-modal attention to generate textual answers.

- **Visual Model:** We use PanDerm, a foundation model pre-trained on dermatological images, as our vision encoder. PanDerm extracts the necessary visual features from skin lesion images which are crucial for accurate reasoning and answering questions about the image.
- **Language Model:** The text input (question) is encoded using BERT (bert-base-uncased), a transformer-based language model. BERT is employed to generate embeddings for the input questions.
- **Decoder:** To produce textual answers, the architecture incorporates a GPT-based decoder. The decoder takes fused multimodal features (from the cross-modal attention) as input and generates natural language answers.

4.2 Training Process

The data preparation process described in Section 3 was used to construct a multimodal dataset consisting of images, associated diagnoses, questions, and answers. The dataset was further split into training, validation, and test subsets. PyTorch's DataLoader was employed to manage batching, shuffling, and efficient data loading to seamlessly handle large datasets during training. Throughout the training process, we kept the vision transformer (PanDerm) frozen to retain its pre-trained feature extraction capabilities, focusing the training on the language model and the cross-modal fusion model.

4.2.1 Loss Function

Training utilized the CrossEntropyLoss from PyTorch to compute the difference between the predicted answers and the ground truth. The loss was calculated at each timestep of the output sequence during text generation and backpropagated to update model parameters.

4.2.2 Optimization

The model's parameters were optimized using the AdamW optimizer. Gradients were computed through backpropagation, which adjusts the model's weights by passing the loss information backward through the network. The training loop included steps to clear gradients, calculate losses, and iteratively optimize the model over multiple epochs.

4.3 Evaluation Metrics

The following metrics was used to evaluate the model's performance:

- **BLEU Score:** Evaluates how closely the predicted answers match the ground truth by measuring the n-gram precision.
- **ROUGE Score:** Captures the similarity between the predicted answers and the ground truth by assessing the overlap of sequences, including precision, recall, and F1-score.

5 Results

For quantitative evaluation, we measure the model's performance using BLEU and ROUGE scores. The evaluation of the model's performance using such scores provides insights into the quality of the generated text compared to the reference text. Below, we discuss the results for each metric separately, focusing on the implications of the scores and the potential reasons behind the observed trends.

The BLEU scores in Table 5 are notably low across all fusion mechanisms and language models. BLEU measures n-gram overlap between generated and reference text, with higher scores indicating better alignment. The low scores suggest insufficient lexical overlap. Consequently, we could observe three important key findings:

- **Short Generated Text:** Manual evaluation revealed that the generated text is much shorter than the reference. BLEU penalizes brevity, as shorter texts have fewer n-gram matches.
- **Lexical Divergence:** The generated text may use different phrasing or vocabulary, even if semantically similar. BLEU focuses on exact word matches, which may not align with the model's output.
- **Model Limitations:** Some fusion mechanisms or language models (e.g., BERT) may struggle to produce text with high lexical overlap.

Fusion Method	**BERT**	**DistilBERT**
Cross-Attention	0.0201	0.0519
UNITER	0.4181	0.6291
ViLT	0.0220	0.6434

Table 4: BLEU - Performance Comparison Across Fusion Mechanisms and Language Models

In contrast, the ROUGE scores in Table 5 are strong, indicating a good match between generated and reference text. ROUGE measures recall, focusing on the overlap of key content and ideas. We summarize our observation as follows:

- **Content Alignment:** The generated text captures the main ideas and key phrases from the reference, even if the wording or length differs.
- **Model Strengths:** Fusion mechanisms (e.g., CrossAttention, UNITER) and language models (e.g., BERT) effectively extract and reproduce core information.
- **Evaluation Focus:** ROUGE is less sensitive to text length, rewarding content overlap even in shorter texts.

Fusion Method	**BERT**	**DistilBERT**
Cross-Attention	0.5547	0.6976
UNITER	0.6369	0.8184
ViLT	0.3672	0.8332

Table 5: ROUGE - Performance Comparison Across Fusion Mechanisms and Language Models

6 Discussion

This study explored the integration of visual and textual modalities for Visual Question Answering (VQA) in dermatology, leveraging the PanDerm foundation model. The primary objective was to augment image-based diagnostic models with natural language capabilities, enabling an interactive system for dermatological analysis.

Our experiments demonstrated that different fusion mechanisms significantly influenced performance. Among them, ViLT achieved the highest BLEU and ROUGE scores, indicating its effectiveness in joint encoding of image and text inputs. However, the relatively lower BLEU scores suggest that the model struggled with generating responses that closely matched reference answers, potentially due to the verbose nature of ground truth responses. The results highlight the challenges of aligning textual and visual representations in a constrained dataset environment.

A key limitation of this study was the difficulty in obtaining a large dermatology-specific VQA dataset. The augmentation of ISIC 2018 with external Q&A data improved multimodal learning, but inherent biases in the dataset could have influenced the model's understanding. Additionally, limited computational resources restricted the ability to fine-tune larger models, which may have impacted overall performance.

Despite these challenges, our approach provides a solid foundation for future work in dermatology VQA. Future improvements could involve fine-tuning lightweight model adaptors (e.g., using LoRA) to optimize computational efficiency, integrating larger multimodal datasets to enhance reasoning capabilities, and incorporating pre-trained large vision-language models for superior contextual understanding.

7 Conclusion

This research presents a significant step towards developing an interactive system for dermatology by integrating a specialized vision model (PanDerm) with language models to enable automated question-answering. Our findings show that multimodal learning enhances the interpretability of dermatological diagnoses while providing user-friendly assistance. The study also emphasizes the critical role of fusion mechanisms in improving performance, highlighting opportunities for further innovation in combining visual and textual data.

However, challenges such as limited dataset availability, model optimization, and computational constraints remain key hurdles. By leveraging advancements in language model fine-tuning and employing advanced dataset augmentation strategies, future iterations of the system can achieve greater accuracy, paving the way for more effective AI-assisted tools that enhance diagnostic precision and support dermatologists.

REFERENCES

Clinic Altruderm. Dermatofibroma removal at the altruderm clinic: 6 of your questions answered, 2024. URL https://altruderm.co.uk/dermatofibroma-removal-at-the-altruderm-clinic-6-of-your-questions-answered/. Accessed: 2024-10-29.

MD Anna Hernández. What is it, causes, signs and symptoms, and more. *Osmosis.org*, 2024. URL https://www.osmosis.org/answers/melanocytic-nevus. Modified: May 15, 2024. Accessed: 2024-10-29.

Marie Bauer, Constantin Seibold, Jens Kleesiek, and Amin Dada. IKIM at MEDIQA-M3G 2024: Multilingual visual question-answering for dermatology through VLM fine-tuning and LLM translations. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman (eds.), *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 439–447, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.clinicalnlp-1.44. URL https://aclanthology.org/2024.clinicalnlp-1.44.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering, 2019. URL https://arxiv.org/abs/1905.13648.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2018. URL https://arxiv.org/abs/1902.03368.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog, 2017. URL https://arxiv.org/abs/1611.08669.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends, 2022. URL https://arxiv.org/abs/2210.09263.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth, 2016. URL https://arxiv.org/abs/1603.09382.

Mreeb. Dermatology question answer dataset for fine-tuning, 2024. URL https://huggingface.co/datasets/Mreeb/Dermatology-Question-Answer-Dataset-For-Fine-Tuning.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161. URL https://doi.org/10.1038/sdata.2018.161.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Gin Tan, Vincent Tang, Aik Beng Ng, David Powell, Paul Bonnington, Simon See, Monika Janda, Victoria Mar, Harald Kittler, H. Peter Soyer, and Zongyuan Ge. A general-purpose multimodal foundation model for dermatology, 2024. URL https://arxiv.org/abs/2410.15038.

W. Yim, A. Ben Abacha, Fu Velvin, Z. Sun, Meliha Yetisgen, Fei Xia, and M. Krallinger. Visual question answering in dermatology (dermvqa), July 2024. URL https://doi.org/10.17605/OSF.IO/72RP3.