

TEAM MIPAL at MEDIQA-M3G 2024: Large VQA Models for Dermatological Diagnosis

Hyeonjin Kim MIN KYU KIM Jae Won Jang
KiYoon Yoo Nojun Kwak*

Seoul National University

{peaceful1, alsrb7000, pert0407, 961230, nojunk}@snu.ac.kr

Abstract

This paper describes the methods used for the NAACL 2024 workshop MEDIQA-M3G shared task (wai Yim et al., 2024a) for generating medical answers from image and query data for skin diseases. MedVInT-Decoder (Zhang et al., 2023b), LLaVA (Liu et al., 2024), and LLaVA-Med (Li et al., 2024) are chosen as base models. Finetuned with the task dataset on the dermatological domain, MedVInT-Decoder achieved a BLEU score of 3.82 during competition, while LLaVA and LLaVA-Med reached 6.98 and 4.62 afterward, respectively.

1 Introduction

The advancement of telecommunication technologies and the increasing demand for healthcare services have accelerated the demand for remote disease diagnosis and treatment. However, existing medical-related multimodal problems have predominantly focused on general diseases or radiology image analysis. In this task, abnormal skin images along with a single conversational query from the patient are provided as inputs. The goal is to utilize a multimodal model to identify the patient's condition and generate appropriate responses from the physician tailored to the patient's situation.

To address this problem scenario, we finetuned three multimodal VQA models, MedVInT-Decoder (Zhang et al., 2023b), LLaVA (Liu et al., 2024), LLaVA-Med (Li et al., 2024). When finetuned with the task dataset (wai Yim et al., 2024b), the BLEU scores of the MedVInT-Decoder, LLaVA, and LLaVA-Med were 3.82, 6.98, 4.62. The results for LLaVA and LLaVA-Med were submitted after the challenge. Since the released train dataset was small, we further explored ways to augment this data. Specifically, we crawled skin disease images online and synthesized query-response pairs using GPT-3.5. However, models

trained on the synthetic data reached a BLEU score lower than 1. The reason for such failure is discussed in Section 5.

2 Related Works

2.1 Visual Question-Answering

Multimodal models that target Visual Question-Answering tasks (VQA) are mostly consisted of a vision encoder, a text encoder and a decoder that decodes the encoded image and text at once. Some models use ViT as the vision encoder (Yu et al., 2022; Chen et al., 2022; Liu et al., 2024) while others employ ConvNet such as ResNet-50 (Wang et al., 2021). The encoded visual features are subsequently processed through the projection layer, where they are transformed into the word embedding space. The language instructions along with the projected image features are concatenated and inputted to the language model decoder to generate the output texts.

2.2 VQA on Medical Domain

MedVInT (Zhang et al., 2023b) uses pretrained ResNet-50 from PMC-CLIP (Lin et al., 2023) as the vision encoder, and PMC-LLaMA (Wu et al., 2023) as the language model. The model is then pretrained on a large dataset for VQA tasks on medical domain. MedVInT comes in two different forms: one uses encoder-based language model and the other uses decoder-based one as the generator. MedVInT-Decoder seems to generate more human-like answers in our experiments and is chosen as a base model.

LLaVA (Liu et al., 2024) employs ViT-L/14 from CLIP (Radford et al., 2021) to encode images, and Vicuna (Chiang et al., 2023) to encode and generate texts. LLaVA-Med (Li et al., 2024) is a LLaVA baseline model finetuned on medical dataset of 178k text queries and 61k images across X-ray, MRI, histology, gross pathology and CT domains.

*Corresponding Author

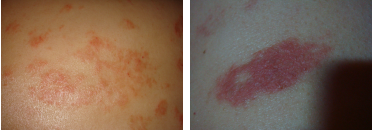
| |
|---|
| <p>encounter_id: ENC00966</p>  <p>query_title: ""</p> <p>query_content_en: "Patient is 46 years old, female. The problem came up at the back of the body after wearing old clothes. Symptom felt: acute itchiness, small amount of discharge leaking from the affected area. It started small, but outburst to patches after washing with warm water. It is not improving after taking Clarityne. Now apply topically Disong camphor thin cream (A Chinese herbal antiinflammatory, antiallergic ointment). Orally taking Ketotifen."</p> <p>Reference: "Based on the medical history and picture, it should be Allergic dermatitis. Use antiallergy treatment."</p> <p>MedVInT workshop: "It is a case of allergic dermatitis."</p> <p>LLaVA 29 diseases: "Scabies. Treatment involves killing mites and eggs with medicated cream or pill."</p> <p>LLaVA 324 diseases: "Malassezia folliculitis. Treat with antifungal agents. Consider oral antifungal medications like ketoconazole."</p> <p>LLaVA workshop*: "Dermatitis"</p> |
|---|

Table 1: Example of generated responses using our pipeline.

3 Method

3.1 Dataset Preprocessing

The chosen base models take a single image-query pair as their input and generate a single response as their output. The task dataset has multiple images for input and multiple possible responses per query. This requires a selection process to match the model input structure.

For each query, we sort the responses by the reliability of the authors, which are determined by their level of expertise. Then, we pair each of the responses with a single image. For example, if a query had 5 images and 3 different answers, three image-answer pairs are created for the query. Since some responses were exactly identical, we remove the duplicated responses to give diversity.

Some authors of the responses are ranked with low reliability score. One can consider removing these from the dataset to improve the validity of the train set. In this paper, we chose not to remove such answers to keep the dataset as large as possible. After the whole process, we obtained 2101 triplets of image, query and response. We train only on the English data.

3.2 Synthetic Data Generation

To augment the limited number of training samples, we attempted to generate synthetic data. Two different datasets were collected from two sources: one with 29 classes of most common diseases (Furue et al., 2011; Li et al., 2023; Zaidi and Lanigan, 2010) and another containing up to 324 dis-

eases (Atlas). The two datasets are denoted as '29 diseases' and '324 diseases' each from below. Associated queries attached to the images are generated with GPT-3.5 to imitate the answers in the given dataset.

3.3 Finetuning MedVInT

We finetune the original MedVInT-Decoder model with the processed dataset. Many training options were tested to find the optimal epochs, batch size, learning rate and early stopping. Then the same tests were executed again with other pretrained language models such as BioMedGPT (Zhang et al., 2023a) and MedAlpaca (Han et al., 2023).

The given training set is not large compared to the model size. The composition of the dataset is also noisy as uncertain and unreliable responses were included as well to maintain the size as much as possible. Such noisiness may have caused the slow convergence and prevented the training loss to converge to a lower number. In most cases, the training loss was over 2.5 which is quite large when considering that the metric used to evaluate loss in the model was Cross Entropy Loss. Training for more than 3 epochs results in overfitting because of the small dataset size.

3.4 Finetuning LLaVA variants

We adopt the pretrained LLaVA model as the baseline and subsequently conduct finetuning of both the projection matrix and the language model using low ranked adaption (LoRA) (Hu et al., 2021). We train the LLaVA for one epoch with a batch size

| Train Data | Model | BLEU | BERT |
|------------|----------------------|------|------|
| - | MedVInT | 0.91 | 0.82 |
| | LLaVA | 0.93 | 0.84 |
| | LLaVA-Med | 1.35 | 0.85 |
| MEDIQA-M3G | MedVInT | 3.82 | 0.87 |
| | LLaVA* | 6.98 | 0.86 |
| | LLaVA-Med* | 4.62 | 0.84 |
| Synthetic | LLaVA (29 diseases) | 0.92 | 0.84 |
| | LLaVA (324 diseases) | 0.98 | 0.85 |

Table 2: BLEU score and BERTscore of models finetuned on different datasets. The scores for LLaVA workshop and LLaVA-Med workshop was attained after the competition. All scores were measured on the final test set.

of 8 and a gradient accumulation step of 16. The LoRA hyperparameter r was set to 128 and α was set to 256. The learning rates were set to $2e-5$ for the projection layer and $2e-4$ following the original configuration. We similarly inputted single image per query as the model was not finetuned on multiple images. Throughout the finetuning process, which spanned 10 epochs, we utilized the validation dataset to select the checkpoint from the epoch with the lowest evaluation loss. We employed the pretrained LLaVA-Med model, which is a version of LLaVA that has been finetuned on a medical dataset (Li et al., 2024). Finetuning was done for 10 epochs with a batch size of 8 and gradient accumulation step of 16.

When training only on the task data, we denote it by "MEDIQA-M3G". The results for training additionally on the synthetic data is denoted by "Synthetic".

4 Results

Examples of the generated responses are provided in Table 1 and summarized results are in Table 2. All models and checkpoints are evaluated using the official test set only.

MedVInT-Decoder shows low BLEU score of 0.91 when the inference is made directly on the test set without any finetuning. During training, MedVInT-Decoder reaches the lowest validation error after around 3 epochs and starts to show signs of overfitting afterwards. Early stopping is introduced to make use of such pattern.

The BLEU score measured with the validation set was the highest when trained with learning rate of $4e-6$, batch size 16 and epochs 10. The train-

*After-challenge submission

ing process stopped early at epoch 3. The BLEU score was 4.48 on the validation set and 3.82 on the test set. Although not impressively high, the increased values prove that actual learning has been conducted.

In case of LLaVA, inferencing with the vanilla model without any finetuning yielded a BLEU score of 0.93 which is similar to that of MedVInT. Upon finetuning the model with 29 diseases set and 324 diseases set, the BLEU scores on the validation dataset showed 0.97 and 4.51, respectively. The latter seemed promising, but did not meet the expected performance when inferenced on the test set.

LLaVA-Med scored a BLEU score 1.35 on zero shot inference. Afterwards, when finetuning on the task data, LLaVA-Med achieved a BLEU score of 3.82 and BERT score 0.84.

5 Discussion

The Necessity of Finetuning Although there already exist models pretrained with medical data, they mostly failed to give high-quality answers without further finetuning on the domain specific data. Two observations were made to explain this phenomenon.

One reason would be the difference of the image domain. Both PMC-VQA and LLaVA-Med training set are composed of professional images such as X-Ray, MRI and CT. The raw photos are not many in these sets, and it becomes even scarcer when limited to skin diseases. Therefore the task images would have been regarded as new and unfamiliar to the pretrained models.

Another possible reason would be the difference in the text domain. MedVInT-Decoder trained with PMC-VQA has learned to give short answers for most of the time. PMC-VQA is consisted of simple yes or no questions, or those that can be answered with one or two vocabularies. This may not have been sufficient when it comes to diagnosis and prescription tasks that require the machine to give long answers. Also, non-negligible amount of texts used for training is structured with certain formats or may be excerpts from academic texts. This is in contrast to the task dataset which is mostly written in casual spoken language.

Low Performance When Trained with Crawled Data Upon observing lower-than-expected performance following finetuning of the model with additional crawled data, efforts were directed towards

enhancing LLaVA’s performance after the competition. Inspired by the impressive performance demonstrated by MedVInT upon finetuning with solely the workshop-provided training data, a similar approach was applied to LLaVA and LLaVA-Med. LLaVA in this method yielded BLEU and BERT scores of 6.98 and 0.86, respectively, representing the highest scores achieved on the test dataset as shown in Table 2. This outcome underscores LLaVA’s superior performance compared to MedVInT and LLaVA-Med. Furthermore, it suggests potential disparities between the dataset synthetically generated by ChatGPT and the real-world data. Lastly, utilizing BLEU and BERT scores as metrics implies that achieving similar linguistic nuances may contribute to superior performance, rather than merely focusing on the accuracy of individual predictions.

6 Conclusion

We propose our submission to the MEDIQA-M3G shared task for generating medical responses to multimodal queries. In our study, three existing models MedVInT, LLaVA and LLaVA-Med are finetuned using the competition dataset along with synthetically generated dataset. Their performance are evaluated using BLEU and BERT score. Our results indicate that utilizing only the task dataset leads to substantial improvements in both models, reaching the BLEU score of 3.82 with MedVInT-Decoder and ranked second in the English section of the workshop. After the competition, LLaVA finetuned with the workshop dataset achieved the highest BLEU score of 6.98 and lastly finetuned LLaVA-Med model with workshop dataset performed a BLEU score of 4.62.

References

Dermatology Atlas. DermatologyAtlas. <https://www.atlasdermatologico.com.br>. Accessed: 2024-04-04.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing

gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.

Masutaka Furue, Souji Yamazaki, Koichi Jimbow, Tetsuya Tsuchida, Masayuki Amagai, Toshihiro Tanaka, Kayoko Matsunaga, Masahiko Muto, Eishin Morita, Masashi Akiyama, et al. 2011. Prevalence of dermatological disorders in japan: a nationwide, cross-sectional, seasonal, multicenter, hospital-based study. *The Journal of dermatology*, 38(4):310–320.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bresssem. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Huanyu Li, Peng Zhang, Zikun Wei, Tian Qian, Yiqi Tang, Kun Hu, Xianqiong Huang, Xinxin Xia, Yishuang Zhang, Haixing Cheng, et al. 2023. Deep skin diseases diagnostic system with dual-channel image and extracted text. *Frontiers in Artificial Intelligence*, 6.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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*.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.
- Zohra Zaidi and Sean W Lanigan. 2010. *Dermatology in clinical practice*. Springer Science & Business Media.
- Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.