VisionQAries at MEDIQA-MAGIC 2024: Small Vision Language Models for Dermatological Diagnosis

Notebook for the ImageCLEF Lab at CLEF 2024

Patrycja Cieplicka[1,*,†], Julia Kłos[1,†] and Maciej Morawski[1,†]

[1]*Independent Researcher, Warsaw, Poland*

Abstract

In this paper, we present our solution for the ImageCLEF 2024 Challenge MEDIQA-MAGIC task, which addresses the problem of Multimodal and Generative Telemedicine (MAGIC) in dermatology. We report the results of directly prompting existing small-scale multimodal models (moondream2 and TinyLLaVA) and evaluate the impact of fine-tuning these models with domain-specific knowledge. The top-performing model, based on the Moondream model and fine-tuned with domain-specific knowledge, achieved a $deltaBLEU_{en}$ score of 8.629. Our findings highlight the effectiveness of these approaches in improving model performance for dermatological applications.

Keywords

small vision language models, image-text-to-text, multimodal, dermatological diagnosis

1. Introduction

The advancement of telecommunication technologies, coupled with the increasing demand for healthcare services and the impact of the recent pandemic, has significantly accelerated the adoption of remote clinical diagnosis and treatment modalities. In addition to synchronous consultations with healthcare providers via telephone or video conferencing, asynchronous communication methods such as emails and chats have demonstrated considerable cost-effectiveness and convenience.

In this context, the Multimodal and Generative Telemedicine (MAGIC) initiative this year concentrated on the field of dermatology. The inputs for the ImageCLEF 2024 Challenge MEDIQA-MAGIC task [1] included text, which provided clinical context and queries, as well as one or more images. The primary challenge was to generate a suitable textual response to the given queries.

In this paper, we present an attempt to solve the task using existing multimodal models. We focus specifically on small-scale models for image-text-to-text generation. Due to their size, those models are suitable for deployment on portable devices, enabling their application in telemedicine scenarios. Device-only models offer a valuable solution for countries with limited or prohibitively expensive internet access.

2. Related work

In the last years, we have witnessed enormous progress in artificial intelligence (AI) mainly in the field of natural language processing (NLP) and computer vision (CV).

In 2017 Ashish et al. [2] presented Transformer architecture, encoder-decoder model, a key element of which is a self-attention mechanism that helps the model focus on more important parts of the sequence, giving more weight to specific words based on their relevance to the task. As the computational

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

patrycja.cieplicka@gmail.com (P. Cieplicka); julia.klos159@gmail.com (J. Kłos); mpfmorawski@gmail.com (M. Morawski)
https://github.com/paatrycjaa (P. Cieplicka); https://github.com/julklos (J. Kłos); https://github.com/mpfmorawski (M. Morawski)

capabilities of GPU (Graphics Processing Unit) increased those two aspects have been crucial in the evolution of Large Language Models (LLMs). One of the initial LLM models is BERT (Bidirectional Encoder Representations from Transformers) was introduced in 2019 [3]. It has achieved state-of-the-art results in numerous NLP benchmarks and tasks. The era of even larger LLMs began in 2020, introducing models like GPT-3 provided by OpenAI [4] that was incredible good at generating human text. Today open-source LLMs such as LLaMA [5], Llama 2 [6], Vicuna [7], and Mistral 7B [8] are also available.

Transformer models have found also numerous valuable applications in the field of computer vision, where so far mainly architectures based on convolutional neural networks have been used. Vision Transformer architecture (ViT) was introduced [9] that treats an image as a sequence of patches and processes these patches using self-attention, allowing the model to capture long-range dependencies and global context. It excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train. In 2021 DALL-E [10] and Stable Diffusion [11] became state-of-the-art text-to-image generation models. CLIP (Contrastive Language-Image Pre-training) [12] model which efficiently learns visual concepts from natural language supervision and can act as a visual encoder and SigLip [13] model, compared to CLIP uses a simpler but computationally cheaper loss function, were also introduced.

In real-world cases, we analyze data from many different sources, so there has been increased interest in visual language models (VLMs). They combine visual and textual information to understand and generate content and are good choices for downstream tasks such as Visual Question Answering (VQA), image captioning as well as medical report generation. Most of the recent VLMs architectures consist of three components 1) a visual representation backbone, 2) a vision-language projector, and 3) a language model [14]. In this approach, a (pre-trained) visual backbone is used to map an input image to the sequence of patch features that are then projected individually into the embedding space of the language model. To better align LMMs with human preferences, recent works, such as LLaVA [15] and InstructBLIP [16], propose fine-tuning VLMs with visual instruction tuning data, which greatly enhances models' zero-shot capabilities.

To adjust VLMs for medical purposes many approaches were presented [17] For example LLaVa-Med model [18] authors proposed a novel curriculum learning method for adapting LLaVA to the biomedical domain using their self-generated biomedical multi-modal instruction-following dataset. Another existing solution is Visual Med-Alpaca [19], a system that connects multiple image captioning models with an LLM, using a classifier to determine if or which biomedical captioning model is responsible for the image. Many of the medical VLM's have focused on radiology images - RaDialog [20] is a novel approach to integrating advanced vision-language models for the generation, interactive modification, and analysis of radiology reports.

However, we need to consider the computation bottleneck usually introduced by LLMs, which are one of the core components of recent VLM architectures. Recent research shows that small-scale LLMs such as Phi-2 [21], TinyLlama [22] and StableLM2 [23] have reached impressive performances while maintaining reasonable compute budgets. Existing small vision language models such as TinyLLaVA [24] or moondream2 [25] show promising results.

3. Proposed solution

Our objective was to evaluate the efficacy of existing multimodal solutions within the provided use case. Given our limited access to extensive computational infrastructure, we elected to concentrate on small-scale Large Multimodal Models (LMMs), specifically Moondream 2 and TinyLLaVA. These models can be successfully run and fine-tuned on a private computer, making them ideal candidates for our research. Fine-tuning was made possible by access to the training dataset provided by the challenge organisers.

3.1. Dataset

The dataset provided by the challenge's organizers consisted of 270 train and 43 validation clinical dermatology textual queries with an associated image, as well as the answers to the queries (after downloading from Reddit) [26]. Finally, the test split comprised 78 clinical dermatology textual queries with an associated image.

3.2. Models

3.2.1. moondream 2

Moondream 2 is an open-source vision-language model [25]. Despite its modest size of 1.86 billion parameters, which is relatively small considering that most effective models typically start at 7 billion parameters, it is specifically designed for efficient operation on devices with limited computational resources, such as Raspberry Pi, edge devices, or mobile phones. Like conventional vision-language models, moondream 2 interprets visual data to generate textual responses based on the provided information. It has demonstrated utility in various applications, including security and retail.

Moondream 2 leverages weights derived from SigLIP and Phi-1.5. Phi-1.5 is a compact language model with 1.3 billion parameters and a transformer-based architecture, trained on the LLAVA training dataset. SigLIP (Sigmoid Loss for Language Image Pre-Training) is a method that facilitates learning by sequentially analyzing images and their captions, enhancing speed and efficacy, especially when handling large datasets. Similar to the CLIP (Contrastive Language-Image Pre-training) model, Moondream 2 differentiates itself by substituting the softmax loss used in CLIP with a simple pairwise sigmoid loss. This alteration ensures improved performance by focusing exclusively on image-text pairs, eliminating the necessity for a global view of all pairwise data within a batch, thereby increasing the efficiency and speed of the training process.

3.2.2. TinyLlava

TinyLLaVA presents a novel framework that integrates smaller-scale Large Language Models (LLMs) with compact vision encoders through an intermediate connector [24]. This framework employs models such as TinyLlama, StableLM-2, and Phi-2, in combination with vision encoders like CLIP and SigLIP. The connector, a two-layer Multi-Layer Perceptron (MLP) with GELU activation, facilitates effective communication between the vision encoders and the small-scale LLMs. This architecture yields a resource-efficient multimodal system without compromising performance quality. The TinyLLaVA framework is inspired by the design principles of LLaVA but provides a generalized implementation.

The tiny-llava-v1-hf model has undergone pretraining on two datasets: the LLaVA dataset and the ShareGPT4V dataset. The pretraining process involved a blend of images and annotations from the following subsets: the 558K LAION-CC-SBU subset, the SAM dataset, and the COCO dataset. The model configuration used in our solution includes TinyLlama/TinyLlama-1.1B-Chat-v1.0 integrated with a CLIP vision encoder.

3.3. Experimental settings

We initiated our investigation by directly prompting the specified models, acknowledging their limited subject-related knowledge. After testing several potential prompts, we narrowed our focus to the two prompts detailed in Table 1 for subsequent experiments. Evaluation metrics calculated during experiments ($deltaBLEU_{en}$ and $BERTScore_{en}$) were the same as used in the challenge [1].

Utilizing evaluation metrics, we observed comparable results, with moondream 2 slightly outperforming TinyLLaVA in terms of $deltaBLEU_{en}$. Given the encouraging performance of these small-scale models, we proceeded to fine-tune Moondream 2 using a train dataset. The fine-tuning process involved extending the initial queries with the provided prompt templates and experimenting with different prompting strategies. The images in the dataset were augmented through a series of transformations, including rotation, horizontal flipping, and colour jitter.

Additionally, we varied the number of epochs for fine-tuning, carefully considering the dataset size to avoid overfitting. This approach optimized model performance and enhanced the model's awareness of dermatology-specific knowledge, making it more proficient in the domain.

Table 1
Prompts templates.

ID	Prompt template
1	USER: <IMAGE> This is additional information about the dermatology issue on the image: <QUERY> What dermatological disease is on the image and how can it be treated?
2	USER <IMAGE> Patient wants to find out what dermatological disease he suffers from. Considering patient additional description: <QUERY> Answer two questions: 1. What dermatological disease is on the image? 2. How can it be treated?

4. Discussion of the results

Table 2 presents the results of all submitted runs on a test set, sorted in descending order of their $deltaBLEU_{en}$ score. The top-performing model uses Moondream 2 architecture, prompt number 1, and is fine-tuned for 10 epochs. The next three models, fined-tuned for fewer epochs, achieve a significant drop in performance compared to the top-performing model. The remaining four models, which do not use fine-tuning, achieve lower $deltaBLEU_{en}$ scores, ranging from 1.614 to 1.404.

The provided results confirm that fine-tuning is an important factor in achieving high $deltaBLEU_{en}$ scores. The top four submissions, which are all run on fine-tuned Moondream 2 models, outperform the remaining four submissions that are run by directly prompting Moondream 2 and TinyLLaVA models. Additionally, the number of epochs used for fine-tuning also appears to have a significant impact on performance. The top-performing model is fine-tuned for the highest number of epochs (10 epochs) and the $deltaBLEU_{en}$ score achieved by it is almost 6 times higher than for the same model without fine-tuning. However, we are aware that there is still much room for improvement.

The results do not show a clear pattern in terms of the impact on the performance of prepared prompt templates. Models when prompted directly without fine-tuning achieved slightly higher results with prompt number 2. However, when the Moondream 2 model was fine-tuned, using prompt number 1 resulted in a significantly higher $deltaBLEU_{en}$ score.

During the qualitative analysis of the top-performing model, over-fitting was observed. Ten out of the seventy-eight answers generated by this model on the test set, presented partially in Table 3, are highly repetitive (identical or almost identical). These answers are almost the same as one of the answers from the train set (also shown partially in Table 3).

5. Conclusions

In conclusion, it was verified that small VLMs can be effectively applied to solve the VQA task in the dermatological domain. The results suggest that fine-tuning small-scale VLMs is an important factor in achieving a high $deltaBLEU_{en}$ score. This process enhances the model's awareness of dermatology-specific knowledge, making the models more proficient in this domain. Utilizing subject-domain knowledge may significantly increase the accuracy of results. However, it can also cause overfitting, potentially leading to the generation of misleading, repetitive answers that may not be relevant to the current context.

Future directions for improvement could include exploring other small-scale multimodal models, refining prompts used for current models, or using more advanced fine-tuning techniques. Another direction of experiments could be exploring large-scale VLMs. Nevertheless, we need to consider that deployment of these models on edge devices could be impossible. Furthermore, while the models show promise in generating relevant responses, clinical validation of models' outputs to ensure their

Table 2
Results on test set sorted from highest to lowest $deltaBLEU_{en}$ score.

Run ID	Model details				Scores	
	Model	Prompt version	Fine-tuned	Num. of epochs	$deltaBLEU_{en}$	$BERTScore_{en}$
6	moondream2	1	YES	10	8.629	0.848
2	moondream2	1	YES	5	4.600	0.843
3	moondream2	2	YES	5	3.420	0.846
4	moondream2	1	YES	2	2.231	0.834
8	moondream2	2	NO	N/A	1.614	0.843
1	moondream2	1	NO	N/A	1.502	0.839
7	tiny-llava-v1-hf	2	NO	N/A	1.487	0.844
5	tiny-llava-v1-hf	1	NO	N/A	1.404	0.839

Table 3
Examples of answers, which were generated by the Moondream 2 model fine-tuned for 10 epochs, that are identical or very similar to an answer from the train dataset.

Set	Encounter id	Answer
Train	11htonk	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11urq4h	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11xqiex	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11q3s37	t is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11gqgnw	It is a case of dermatitis due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	125d5do	Most probably it is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11hfupz	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11x2q0q	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11rxhph	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	1248w9o	It is a case of eczema due to dry skin (xerosis). [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed.
Test	11tjsat	It is a case of eczema due to dry skin (xerosis), contact, allergy, or atopic eczema. [...] Use topical steroid cream twice daily for 2-4 weeks and oral antihistamines if needed. [...] as a possible diagnosis.

alignment with clinical guidelines and standards is essential. Collaboration with dermatology experts and clinicians could be crucial in real-world scenarios.

It is also worth mentioning that while these models can provide support during telemedicine consultations or prescreening, they should not fully substitute an appointment with a specialist.

References

[1] W. Yim, A. Ben Abacha, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology, in: CLEF 2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin,

Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, D. Amodei, Language Models are Few-Shot Learners, in: NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates, Inc., 2020, p. 1877–1901.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, ArXiv abs/2302.13971 (2023).

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2020).

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8821–8831.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2021. arXiv:2112.10752.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[13] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986.

[14] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, D. Sadigh, Prismatic vlms: Investigating the design space of visually-conditioned language models, arXiv preprint arXiv:2402.07865 (2024).

[15] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, Llava-next: Improved reasoning, ocr, and world knowledge, 2024. URL: https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[16] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, Advances in Neural Information Processing Systems 36 (2024).

[17] I. Hartsock, G. Rasool, Vision-language models for medical report generation and visual question answering: A review, arXiv preprint arXiv:2403.02469 (2024).

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

[19] C. Shu, B. Chen, F. Liu, Z. Fu, E. Shareghi, N. Collier, Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities, 2023.

[20] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, M. Keicher, Radialog: A large vision-language model for radiology report generation and conversational assistance, arXiv preprint arXiv:2311.18681 (2023).

[21] M. Javaheripi, S. Bubeck, M. Abdin, J. Aneja, S. Bubeck, C. C. T. Mendes, W. Chen, A. Del Giorno, R. Eldan, S. Gopi, et al., Phi-2: The surprising power of small language models, Microsoft Research

Blog (2023).

[22] P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: An open-source small language model, arXiv preprint arXiv:2401.02385 (2024).

[23] M. Bellagente, J. Tow, D. Mahan, D. Phung, M. Zhuravinskyi, R. Adithyan, J. Baicoianu, B. Brooks, N. Cooper, A. Datta, et al., Stable lm 2 1.6 b technical report, arXiv preprint arXiv:2402.17834 (2024).

[24] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, L. Huang, Tinyllava: A framework of small-scale large multimodal models, 2024. arXiv:2402.14289.

[25] vikhyat, Moondream, https://github.com/vikhyat/moondream, 2024.

[26] W. Yim, Y. Fu, Z. Sun, A. Ben Abacha, M. Yetisgen, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, CoRR (2024).

A. Online Resources

The source code of our approach is available via GitHub.