2024 5th International Conference on Computer Engineering and Application (ICCEA) | 979-8-3503-8677-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCEA62105.2024.10603938

Enhanced Qwen-VL 7B Model via Instruction Finetuning on Chinese Medical Dataset

Jianping Luo
TCL Corporate Research (Hong Kong)
Hong Kong, China
luojp@tcl.com

Hanyi Yu
University of Southern California
Los Angeles, California
hanyiyu@usc.edu

Cong Tan
Northeastern University
Shen Yang, China
a13823311216@126.com

Haitao Yu*
Dalian University of Foreign Languages
Dalian, China
yumomo1078515747@qq.com

Jianping Luo and Hanyi Yu are first co-authors.

***Abstract*—The integration of artificial intelligence (AI) into healthcare is revolutionizing medical information access and professional advice delivery. This study focuses on the development of a medical-specific question-answering model leveraging the Qwen-VL 7B model, an advanced Large Vision-Language Model (LVLM), to enhance the understanding and generation of medical texts. The Qwen-VL 7B model's capabilities in both language and visual comprehension make it an ideal candidate for medical question-answering systems, which require a deep grasp of medical knowledge and language. The research objectives include adapting the Qwen-VL 7B model to medical terminology, utilizing its visual understanding for medical imaging analysis, optimizing question-answering systems for medical scenarios, and evaluating the performance of these models. To achieve these goals, we employed methodologies such as dataset construction, model fine-tuning, and user studies. Results showed that the Qwen-VL-Mediacal model achieved a Rouge-1 score of 0.6147, indicating its potential for medical applications. However, challenges remain, such as understanding complex scenes and abstract concepts. Future research will aim to improve the model's adaptability and reasoning capabilities.**

Keywords; Qwen-VL 7B Model ; Medical Question-Answering Systems; Large Vision-Language Model

I. INTRODUCTION

In the contemporary era of information technology, the field of healthcare is undergoing a transformation unprecedented in its history. The rapid advancement of artificial intelligence (AI) has led to the emergence of medical question-answering systems as innovative intelligent tools, altering the way individuals access medical information and professional advice. However, the complexity and specificity of the medical domain necessitate that these systems possess not only robust language comprehension capabilities but also a deep understanding of medical knowledge. Against this backdrop, question-answering models based on Large Language Models (LLMs) have become a focal point of research[1-5].

The Qwen-VL 7B model, as an advanced Large Vision-Language Model (LVLM) [6], has demonstrated exceptional performance in understanding and generating text. Its ability to process not only textual information but also visual content presents immense potential for application in the medical question-answering domain. The Qwen-VL 7B model was designed to enhance machines' capabilities in both visual and language understanding tasks, aligning perfectly with the requirements of medical question-answering systems.

The purpose of this study is to develop a medical-specific question-answering model based on the Qwen-VL 7B model. By integrating medical domain expertise with the powerful language processing capabilities of the Qwen-VL 7B model, we aim to construct a system that can accurately understand users' medical-related inquiries and provide professional responses. This not only enhances the efficiency of medical information acquisition but also assists physicians to a certain extent in diagnostic and treatment decision-making. The research questions are primarily focused on the following aspects:

1. How to adapt the Qwen-VL 7B model to the medical field to enable understanding and processing of medical jargon and concepts.

2. Utilizing the visual understanding capabilities of the Qwen-VL 7B model to process and analyze medical image data, such as X-rays and MRI scans.

3. Designing and optimizing question-answering systems to ensure their accuracy, reliability, and user-friendliness in medical scenarios.

4. Strategies for evaluating and validating the performance of medical question-answering models to ensure their effectiveness and safety in practical applications.

To achieve these objectives, we will employ a range of research methodologies, including the construction of datasets, model fine-tuning, performance assessment, and user studies. We believe that through these research efforts, we can provide the medical field with a powerful intelligent question-answering assistant, thereby contributing to the advancement of healthcare.

This research will contribute to the literature by providing insights into the adaptation of advanced AI models to specialized domains, particularly in the context of healthcare. The findings will be of interest to researchers, healthcare professionals, and AI developers, as they seek to harness the potential of AI to improve patient care and medical outcomes.

II. METHODS

A. Qwen-VL-Mediacal

Despite the significant progress made by multimodal or visual question-answering pre-trained models in handling cross-modal information, the specificity of the medical imaging question-answering field lies in its deep reliance on professional knowledge and precise interpretation of image details. These requirements pose higher challenges in both semantic understanding and visual recognition. Therefore, to better serve medical imaging question-answering tasks, it is necessary to develop a pre-trained model specifically tailored to this domain.

In this study, we trained the Qwen-VL 7B model, which is focused on processing medical imaging data. The model's size reaches 9.07 GB, providing it with substantial computational power for handling large-scale medical imaging datasets. After comparative experiments with other models such as LLava and VisualGlm, we found that Qwen-VL 7B performs more outstandingly on medical imaging question-answering tasks. This achievement not only brings new technological breakthrough. Table 1 give the parameters of Qwen-VL 7B.

TABLE I. QWEN-VL 7B MODEL PARAMETERS

Vision Encoderl	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B

In this study, we used the Qwen-VL 7B model to train the medical image question answering data. Firstly, we integrated the dataset containing the X-ray image questions and answers, set the maximum length of the model to 1800, and performed specific preprocessing on the data Then, the batch size was adjusted to 2, the gradient accumulation strategy was used for 8 times accumulation, and the learning rate was set to 1e-5. In addition, AdamW was selected as the optimizer, and cosine distance was used as the learning plan. After training, the model reached 0.6147 on the Rouge-1 evaluation index, and the loss rate was reduced to about 0.75 We named this model Qwen-VL-Mediacal.

B. Data Process

In this study, we have constructed a multimodal dataset tailored for medical imaging diagnosis, aiming to enhance the capabilities of medical imaging analysis models in understanding and generating relevant diagnostic reports. The construction process of the dataset involved several key steps:

1. Data Source:

Our dataset was sourced from the National Library of Medicine's (NLM) Open-i platform [7]. This platform provides an open-access biomedical image search engine containing over 3.7 million images from approximately 1.2 million PubMed Central® articles, 7,470 chest X-ray images with radiology reports, 67,517 images from the NLM's historical medical collection, and 2,064 orthopedic illustrations.

2. Data Preprocessing:

After acquiring the image data, we conducted a detailed analysis of the diagnostic reports. We utilized the GPT-4 [8] model to generate a series of questions and answers based on each image's diagnostic report. This process not only enriched the dataset's diversity but also increased the model's potential for application in real-world medical scenarios.

3. Question and Answer Generation:

The GPT-4 model considered the accuracy and professionalism of medical terminology when generating questions and answers. The model was trained to understand complex medical concepts and pose questions closely related to the image content. Additionally, the generated answers aimed to provide accurate diagnostic information to support clinical decision-making.

4. Dataset Structure:

Our multimodal dataset comprises three parts: images, questions, and answers. Each image is accompanied by one or more questions, along with corresponding answers. The generation process of questions and answers ensures the dataset's practicality, making it suitable for training and evaluating medical imaging analysis models.

5. Dataset Application:

The medical imaging dataset constructed in this study is designed to support the development of medical imaging analysis models, especially in the fusion applications of Natural Language Processing (NLP) and Computer Vision (CV). The dataset leverages the multimodal capabilities of the Qwen-VL model, which can process image, text, and bounding box inputs and output text and bounding boxes. Qwen-VL excels in multimodal tasks, supports multilingual dialogue, and possesses fine-grained image understanding capabilities.

Through this dataset, we aim to train models that can accurately understand and generate medical diagnostic reports, enhancing the accuracy of medical imaging analysis. The high-resolution input feature of Qwen-VL allows the model to capture more details when processing medical images, thereby providing more precise services in the medical field. The application of this dataset will advance the development of medical imaging analysis technology, offering robust assistance for clinical diagnosis.

C. Pretrained VL-Model

Qwen-VL is a Large Vision Language Model (LVLM) designed to process and understand images and the text information associated with them. The network architecture of Qwen-VL consists of three main components:

- Large Language Model:

The foundation of Qwen-VL is a large language model, which is initialized with pre-trained weights derived from the Qwen-7B [10] model. Qwen-7B is a powerful language model capable of understanding and generating natural language text.

- Visual Encoder:

The visual encoder adopts the Vision Transformer (ViT) [9] architecture, which is a vision processing architecture based on the Transformer model. ViT generates a set of image features by dividing the input image into patches and processing these patches with a certain stride. The visual encoder adjusts the

input image to a specific resolution during both training and inference stages.

- Position-aware Vision-Language Adapter:

To address the efficiency issues that may arise from long sequences of image features, Qwen-VL introduces a vision-language adapter that compresses image features. The adapter contains a single-layer cross-attention module that uses a set of trainable vectors (embeddings) as query vectors and the image features generated by the visual encoder as keys for cross-attention operations. This mechanism compresses the visual feature sequence to a fixed length of 256 for subsequent processing. To preserve the position information required for fine-grained image understanding, the adapter incorporates 2D absolute position encoding into the cross-attention mechanism.

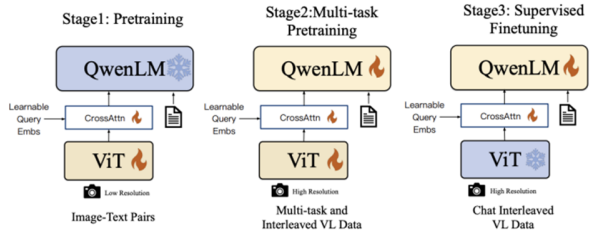

Figure 1. The training pipeline of the Qwen-VL series

The design of Qwen-VL enables it to handle multimodal inputs, including images and text, which makes it excel in tasks such as image captioning, visual question answering, and text localization. By combining the processing capabilities of both vision and language, Qwen-VL provides robust support for understanding complex scenes and performing related tasks.

The training process of this model is divided into three stages: two pre-training phases and a final fine-tuning phase for instruction following, as depicted in Figure 1.

In the first pre-training phase, a large-scale, weakly labeled web-crawled image-text pair dataset was primarily used. This dataset was composed of multiple publicly accessible sources and some internal data. During the data cleaning process, certain specific patterns of data were removed. The original dataset contained approximately 5 billion image-text pairs, which was reduced to 1.4 billion after cleaning, with English data accounting for 77.3% and Chinese data for 22.7%. In this stage, the large language model was frozen, and only the visual encoder and vision-language adapter were optimized. The input images were adjusted to a resolution of 224x224, and the training objective was to minimize the cross-entropy of text labels. The maximum learning rate was set to 2e-4, with a batch size of 30720, and the entire pre-training phase lasted for 50,000 steps, processing about 1.5 billion image-text samples.

In the second multi-task pre-training phase, high-quality, fine-grained visual-language (VL) annotated data was introduced, and the input resolution was increased while processing interlaced image-text data. This stage trained seven tasks simultaneously. To maintain the language model's capabilities, text generation utilized an internally collected corpus. The image captioning data was the same as in Table 2 but with fewer samples and excluding LAION-COCO. Visual question answering (VQA) tasks utilized various public datasets. Additionally, training samples were constructed from datasets such as GRIT, Visual Genome, and RefCOCO to improve text-oriented tasks. To enhance the model's multilingual and multi-image understanding capabilities, additional dialogue datasets were built through manual annotations, model generation, and strategic linking. During this stage, the input resolution for the visual encoder was increased from 224x224 to 448x448 to reduce information loss caused by image downsampling. At the same time, the large language model was unlocked, and the entire model was trained.

In the third stage, the instruction fine-tuning phase, the pre-trained Qwen-VL model was fine-tuned with instructions to enhance its instruction following and conversational abilities, resulting in the interactive Qwen-VL-Chat model. The multimodal instruction fine-tuning data mainly came from image description data or dialogue data generated by LLM self-instructions, which typically involved single-image dialogues and reasoning limited to image content understanding. During the training process, a mix of multimodal and pure text dialogue data was used to ensure the model's versatility in conversational abilities. The instruction fine-tuning dataset consisted of 350k samples. In this stage, the visual encoder was frozen, and the language model and adapter modules were optimized.

D. Applicant

- Text Input:

The model first breaks down the input text into smaller units called tokens, which can be words, subwords, or characters. These tokens are then converted into vector form through an embedding layer, allowing the model to process textual information.

- Image Input:

Images are first processed by the visual encoder and adapter to generate a fixed-length sequence of image features. To distinguish between image feature inputs and text feature inputs, the model adds two special token symbols (<img> and </img>) at the beginning and end of the image feature sequence, indicating the start and end of the image content.

- Bounding Box Input and Output:

To enhance the model's ability for fine-grained visual understanding and localization, Qwen-VL uses data in the form of region descriptions, questions, and detections during the training process. Unlike conventional image-text description or question tasks, this task requires the model to accurately understand and generate region descriptions in a specified format.

For a given bounding box, normalization processing is applied (ranging within [0, 1000]), and it is converted into a specific string format, such as: "(Xtopleft, Ytopleft), (Xbottomright, Ybottomright)".

This string is tokenized as text, without the need for additional positional vocabulary. To differentiate the detection string from regular text strings, two special token symbols (<box> and </box>) are added at the beginning and end of the bounding box string.

Furthermore, to appropriately associate the bounding box with its corresponding descriptive text or sentence, another set of special token symbols (<ref> and </ref>) is introduced to mark the content referenced by the bounding box.

This approach enables Qwen-VL to more accurately understand and generate text descriptions related to image content, especially in tasks requiring fine-grained visual understanding, such as the localization and description of image regions. Through this method, the model can better perform multimodal tasks like visual question answering and image description generation.

- Text Output:

The model recursively predicts the next most likely token based on the current context, and this process continues until it encounters a specific end token like "<eos>" or reaches a preset maximum output length. At each prediction step, the model outputs a probability distribution representing the likelihood of each possible token in the vocabulary. Then, through decoding strategies such as greedy decoding, beam search, or random sampling, a token is selected from the probability distribution. The selected token is recombined into a coherent text output. In specific application scenarios, such as image description tasks, the generated text will also be related to the image content, providing a description or explanation of the image. Figure 2 is an application demo.

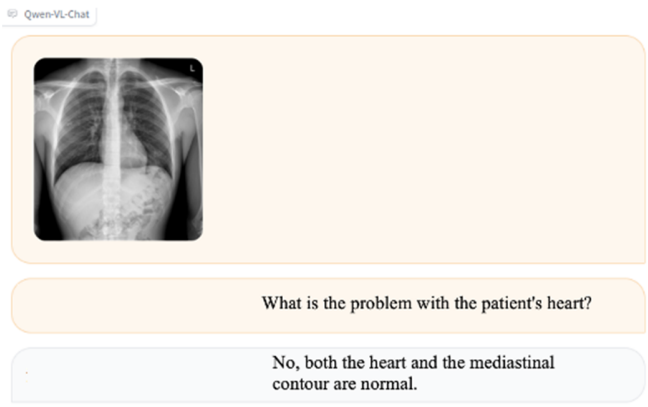

Figure 2. Xray Image Chinese QA demo

III. Results and Analysis

A. Data Set

Based on the OpenNI framework, the author has constructed a visual question-answering (VQA) dataset for the medical field, which has been organized into JSON files. After cleaning, the dataset comprises a total of 7,000 X-ray images, divided into training and testing sets using a five-fold cross-validation method. Each patient may have multiple X-ray images, and questions about their medical conditions are generated from their medical records, along with answers. The data is then formatted according to the requirements of Qwen-VL, primarily by adding a conversational format. This effort ensures the comprehensiveness and stability of the dataset, providing a valuable resource for subsequent research and applications in medical imaging VQA...

B. Training

In our constructed medical image question-answering dataset, we fine-tuned the Qwen-VL model for seq2seq tasks. We set the sequence length to 1800 and designated a Qwen-VL mode data processor for preprocessing medical images and related textual data. During training, we used a batch size of 642 and continued until the loss value decreased to approximately 0.75. The Rouge Loss evaluation indicated that the quality of the model-generated Q&A had reached a reasonable level, culminating in a stable Qwen-VL-Chat model. In this model, users can guide the generation of diagnostic reports or answers by specifying particular types of medical images, along with key medical terms and symptom descriptions, directly influencing the model's output through these key details.

QLoRA is an efficient fine-tuning method allowing for the fine-tuning of models with up to 6.5 billion parameters on a single 48GB GPU, maintaining performance for 16-byte fine-tuning tasks. This method fine-tunes through a frozen int4 quantized pre-trained language model, backpropagating gradients to low-rank adapters, LoRA, with innovations in memory-saving techniques including 4-bit NormalFloat (NF4) data types, dual quantization, and paging optimizers.

ROUGE is an automatic text summarization evaluation metric, primarily measuring the similarity between generated summaries and reference summaries. ROUGE metrics include ROUGE-1, ROUGE-2, and ROUGE-L, focusing on different aspects of textual similarity:

- ROUGE-1 measures word-level overlap between the generated and reference summaries.

- ROUGE-2 considers the overlap of consecutive word pairs.

- ROUGE-L assesses sequence-level similarity based on the Longest Common Subsequence (LCS).

These metrics are commonly used to evaluate text summarization systems, especially in news and conference summarization tasks, with higher ROUGE scores indicating closer similarity to human-generated reference summaries. The loss curve is shown in figure 3.

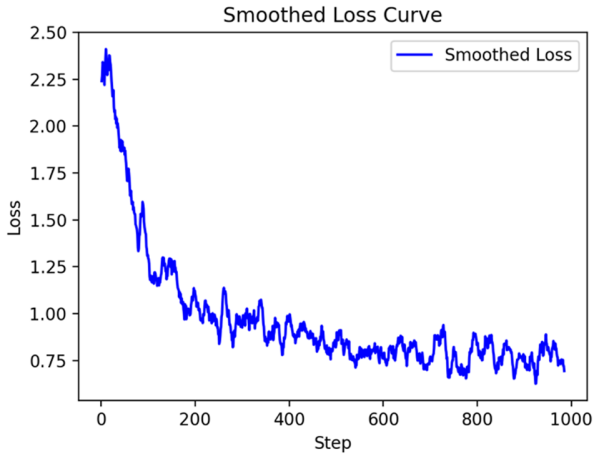

Figure 3: Smoothed loss curve

C. Analysis

In Table 2, we compared the performance of three visual question-answering models: LLava, VisualGlm, and Qwen-ql. The Qwen-ql model outperformed LLava and VisualGlm across all evaluated ROUGE metrics, showing significant improvement, particularly in ROUGE-1. However, the study also highlighted limitations in current visual question-answering models, such as their limited ability to understand complex scenes with obstructions and challenges in processing abstract concepts or questions requiring deep reasoning, which restricts their broader application.

TABLE II. MODEL RESULTS

Models	Rouge-1	Rouge-2	Rouge-L
LLava	0.5801	0.3415	0.5521
VisualGlm	0.5816	0.3440	0.5559
Qwen-ql	0.6147	0.3535	0.5796

IV. CONCLUSION

This study successfully developed a medical question-and-answer system based on the Qwen-VL 7B model, which exhibits outstanding performance in understanding and generating medical-related texts. By integrating medical expertise with advanced large language models, we have constructed a system capable of accurately comprehending users' medical questions and providing professional answers. The Qwen-VL 7B model's performance in medical imaging question-and-answer tasks is particularly noteworthy. Its high scores on the Rouge-1, Rouge-2, and Rouge-L evaluation metrics demonstrate its potential application in the medical field. In terms of dataset construction, we obtained a wealth of medical imaging resources from the Open-i platform of the United States National Library of Medicine. Combined with the GPT-4 model, we generated a diverse set of questions and answers, providing high-quality data for model training. Furthermore, we fine-tuned the Qwen-VL model to better adapt to medical imaging question-and-answer tasks and enhanced its conversational abilities through instruction fine-tuning. Despite significant achievements, the study also identified some challenges. For instance, the model's ability to understand images with complex backgrounds and obstructions is limited, which may affect its application in real medical scenarios. Additionally, the model faces difficulties in dealing with abstract concepts that require deep reasoning. To overcome these challenges, future research could explore ways to improve the model's adaptability to complex scenes and enhance its reasoning and understanding capabilities.

REFERENCE

[1] Salaberria A, Azkune G, de Lacalle O L, et al. Image captioning for effective use of language models in knowledge-based visual question answering[J]. Expert Systems with Applications, 2023, 212: 118669.

[2] Guo J, Li J, Li D, et al. From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10867-10877

[3] Yu S, Cho J, Yadav P, et al. Self-chained image-language model for video localization and question answering[J]. Advances in Neural Information Processing Systems, 2024, 36.

[4] Sharma H, Jalal A S. Image captioning improved visual question answering[J]. Multimedia tools and applications, 2022, 81(24): 34775-34796.

[5] Yang A, Miech A, Sivic J, et al. Zero-shot video question answering via frozen bidirectional language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 124-141.

[6] Bai J, Bai S, Yang S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities[J]. arXiv preprint arXiv:2308.12966, 2023.

[7] Mehnert R B. The National Library of Medicine[J]. Western Journal of Medicine, 1976, 124(4): 346.

[8] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[9] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[10] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.