







Leveraging Large Vision-Language Models for Visual Question Answering in VizWiz Grand Challenge

Bao-Hiep Le* , Trong-Hieu Nguyen-Mau* , Dang-Khoa Nguyen-Vu ,
Vinh-Phat Ho-Ngoc , Hai-Dang Nguyen , Minh-Triet Tran 
University of Science, VNU-HCM

Vietnam National University, Ho Chi Minh City, Vietnam

lbhiep20@apcs.fitus.edu.vn, nmthieu@selab.hcmus.edu.vn, nvdkhoa20@apcs.fitus.edu.vn,
hmvphat21@apcs.fitus.edu.vn, nh dang@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

Abstract

This paper introduces an ensemble-based approach for Visual Question Answering aimed at enhancing accessibility for individuals with vision impairments. By leveraging large vision-language models, our method improves VQA system performance for answering visual questions and image classification. Through rigorous experimentation, we demonstrate the effectiveness of our approach, contributing to advancements in assistive technologies and computer vision research. On the 2024 VizWiz VQA Challenge, we achieve an accuracy of 75.54.

1. Introduction

The Visual Question Answering (VQA)[1] task involves developing algorithms that enable computers to understand and answer questions about visual content. It requires systems to interpret both the image and the natural language question about the content of the image to generate an accurate response. VQA has diverse applications, from assisting the visually impaired community to improving content-based image retrieval systems.

The VizWiz-VQA[4] dataset originates from a natural visual question-answering setting, in which each blind person takes a photo and records the question, talking about it. The challenge is that the photo can be deficient in quality or lacking information, making the question unanswerable.

In previous VizWiz VQA Challenges, the common architecture consisted of a vision encoder, a text encoder, and a text decoder. The vision encoder will encode the image, the text encoder will encode the input question, and the text decoder will output the answer to the question.

The advance of large vision-language models ([2], [3], [7], [8]) has brought a new approach to the VQA task. Using prompting, we can leverage their impressive capabilities in perceiving texts and images to answer visual questions.

*These authors contributed equally to this work

2. Methodology

The overall architecture of our solution is depicted in Figure 2. Our approach is based on prompting the large vision language models to output answers in a suitable format for the challenge. Figure 1 shows an example of model interaction.

We experiment with various vision-language models, including Qwen-VL-Chat [2], LLaVA-1.5 [7], LLaVA-1.6 [8], and InternVL-Chat-V1.2 [3] models. Each model is fine-tuned on the VizWiz dataset using parameter-efficient fine-tuning [5] techniques such as low-rank adaption (LoRA) [6] and weight-decomposed low-rank adaption (DoRA) [9]. Additionally, to boost the accuracy on number-related questions (e.g., counting objects), we employ a subset of the VQAv2 datasets consisting of similar questions.

Each model is fine-tuned with various configurations (e.g., model size, LoRA/DoRA rank). Empirically, we find that without ensembling, LLaVA-1.6 performs best, followed by LLaVA-1.5, Qwen-VL-Chat, and InternVL-Chat-V1.2 (see Table 1). Finally, we employ an ensemble approach to select the most frequent output of 33 fine-tuned models, favoring better-performing ones: 16 LLaVA-1.6





IMAGE: 

PROMPT: Can you tell me what this medicine is please? When the provided information is insufficient, respond with “unanswerable”. Answer the question using a single word or phrase.

MODEL: night time

Figure 1. An example of prompting a vision-language model to answer a visual question

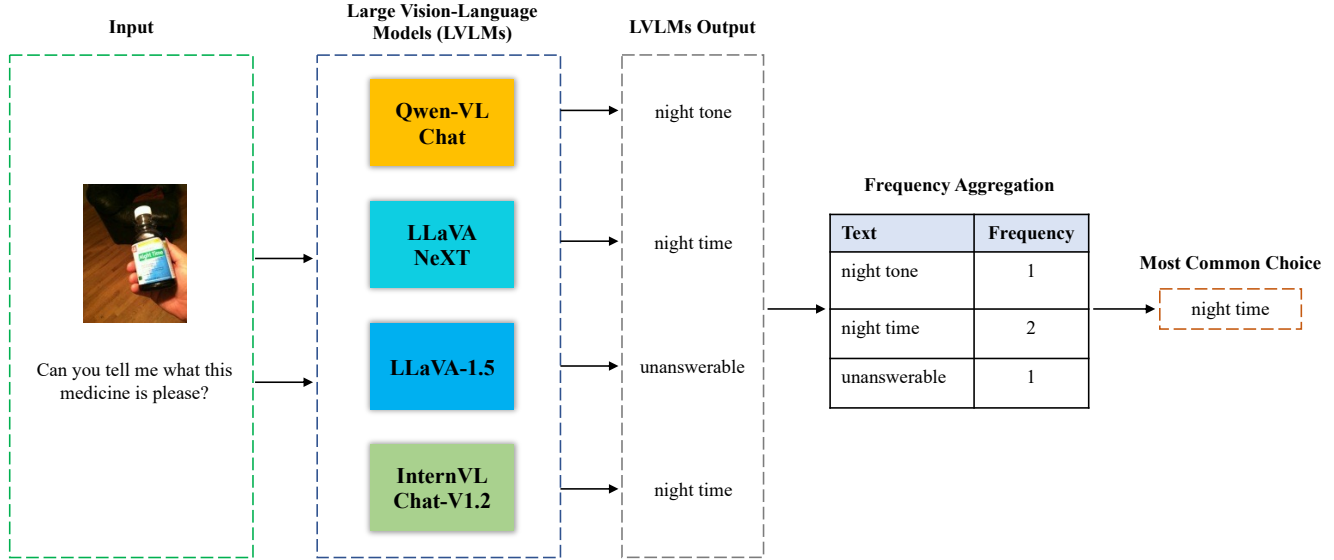


Figure 2. Overall architecture of our proposed solution

models, 9 LLaVA-1.5 models, 6 Qwen-VL-Chat models, and 2 InternVL-Chat-V1.2. Increasing the number of models beyond 33 does not lead to a significant improvement in overall accuracy.

3. Results

Our ensemble model achieved an overall accuracy of 76.00 and **75.54** on the VizWiz test-dev and test-standard splits, respectively. Further evaluation aimed at removing underperforming models may improve overall results.

Model	Overall Accuracy
Qwen-VL-Chat [2]	69.25
LLaVA-1.5 [7]	70.71
LLaVA-1.6 [8]	72.76
InternVL-Chat-V1.2 [3]	66.99
Ensemble (4 models)	73.77
Ensemble (21 models)	75.53
Ensemble (33 models)	76.00

Table 1. Overall accuracy of various models on the VizWiz test-dev split. The best accuracy of all training configurations is reported for each single model.

Acknowledgement: This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We thank University of Science, VNU-HCM, for providing access to the DGX A100 server used in this project.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh.

Vqa: Visual question answering. In *Proceedings of ICCV*, 2015.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[4] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of CVPR*, pages 3608–3617, 2018.

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

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

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

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

[9] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024.