Vision - Large Language Model For Vietnamese Visual Question Answering

Hoang Dinh Quang

University of Information Technology - UIT

What ?

This research focuses on fine-tuning **Qwen2VL**, a **Vision-Language Large Model (VLLM)** that supports Vietnamese, to improve performance on the **Vietnamese Visual Question Answering (VQA)** task. We leverage **LoRA (Low-Rank Adaptation)** to efficiently fine-tune the model on a Vietnamese VQA dataset. The goal is to enhance the model's ability to understand and generate accurate answers based on images and text prompts in Vietnamese.

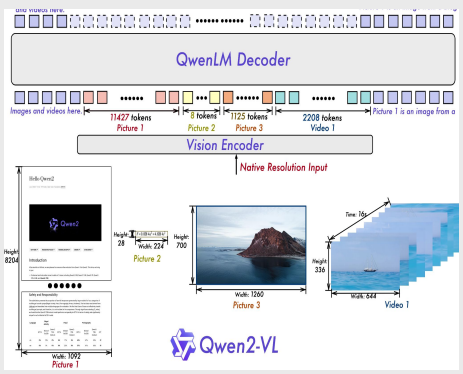
Why ?

Current Vision-Language Large Models (VLLMs) still have limitations in **Vietnamese Visual Question Answering (VQA)** due to insufficient optimization for Vietnamese. These models often struggle with linguistic nuances, leading to less accurate answers. By fine-tuning **Qwen2VL** with **LoRA**, this research enhances the model's ability to process Vietnamese text and visual inputs, improving applications in education, accessibility, and automated content understanding.

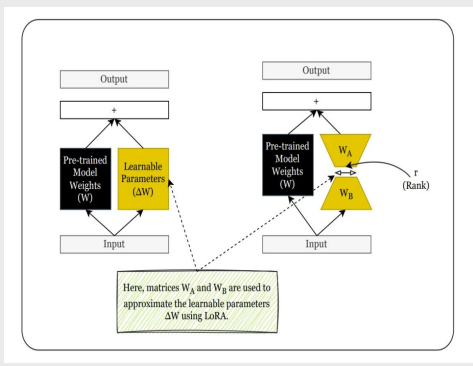
Overview

Qwen2VL → Fine Tuning With LoRA → Evaluate

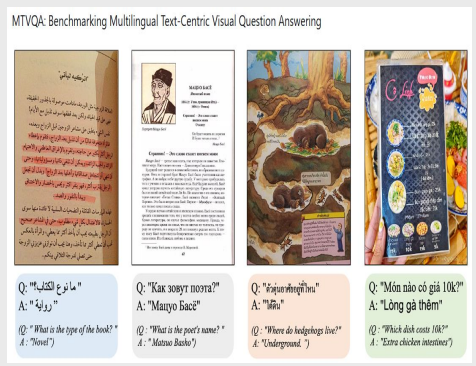
Qwen2VL architecture



LoRA method



MTQVA Benchmark



Description

1. Qwen2VL

- Qwen2VL is a Vision-Language Model (VLM) developed by Alibaba, designed for multimodal tasks such as Visual Question Answering (VQA), image captioning, and text-image understanding. It supports multiple languages, including Vietnamese, making it suitable for diverse applications.
- Built on a transformer-based architecture, Qwen2VL leverages large-scale vision-language pretraining and advanced fine-tuning techniques to enhance text-image reasoning. It excels in extracting meaningful insights from images and generating accurate, context-aware responses.

2. Fine Tuning With LoRA

- The model is fine-tuned using a Vietnamese VQA dataset, optimizing parameters to improve comprehension of image-text interactions in Vietnamese.
- LoRA (Low-Rank Adaptation) is applied to efficiently fine-tune Qwen2VL by injecting low-rank trainable adapters into pre-trained weight matrices, reducing computational cost while maintaining model performance.
- By leveraging LoRA, the fine-tuning process requires fewer GPU resources, enabling faster training and better scalability without modifying the original model weights.

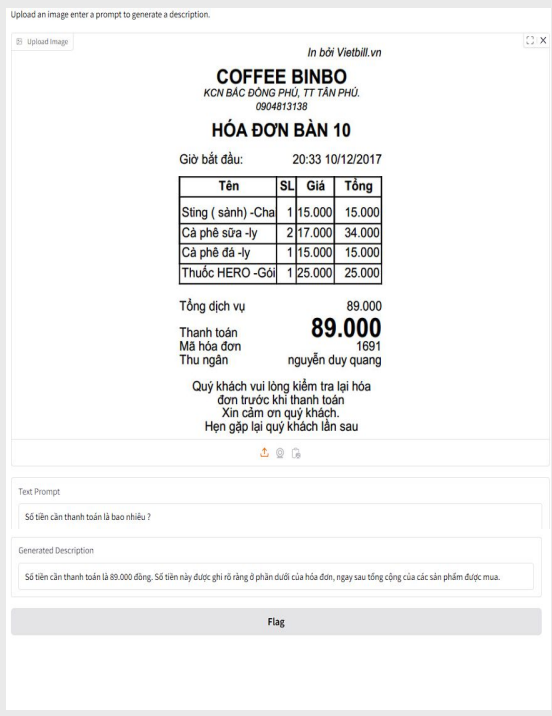


Figure 1. Demo of Inference Results from the Trained Qwen2VL Model on VQA Tasks

3. Evaluate

- Use VQA evaluation metrics (Accuracy, BLEU, METEOR, etc.) on the MTVQA benchmark dataset
- Compare the results with the original Qwen2VL model and other VQA models
- Analyze cases where the model provides incorrect answers to gain insights and propose improvements.

Hoang Dinh Quang – University of Information Technology
TEL : 0945356057 Email : quanghd.18@grad.uit.edu.vn