Amazon ML Challenge 2024
Team : CTRL + ALT + DEV
Amrita Vishwa Vidyapeetham, Coimbatore

Image-Based Entity Value Extraction with Qwen2-VL Model

Team Members :
- Vishal S
- Saran Dharshan S.P
- Amritha Nandini
- Suriya KP

Approach:

Extracting entity values from product images are crucial for sectors like healthcare, e-commerce, and content moderation, where product details such as weight, dimensions, wattage, and voltage are vital for accurate listings. Our approach focused on utilizing the **Qwen2-VL-2B-Instruct model** [1] [2], making use of its vision-language capabilities to interpret both visual and textual information within images.

Image Pre-processing: We resized images, ensuring they fit within the model's input constraints and **enhanced the contrast** of the images using PIL's ImageEnhance to optimize the input quality for better text generation.

Prompt Engineering: In our approach, we crafted targeted prompts based on the product entity, such as "Identify the height of the product" or "Determine the wattage of the product." These prompts were **dynamically generated based on the entity name**, ensuring that the model's focus was directed towards extracting relevant information from the product image and its context. If the required entity is not present in the image, the model must return an empty string.

Model Inference: The preprocessed images, along with their respective prompts, were passed into the Qwen-2-VL-2B-Instruct model, which generated descriptive text. We used a conditional generation approach to extract relevant information, such as weight, dimensions, or voltage. The model was run on **A6000 GPUs** to ensure high computational efficiency. We processed the images in batches of 100, saving the generated text to a CSV file after each batch to ensure incremental progress and prevent data loss.

Post Processing: After obtaining the descriptive text output from the model, we applied post processing techniques to convert the generated text into the desired format. The model's output typically contained additional context, such as full sentences with product descriptions, measurements, or specifications. To refine this into the required format, we extracted only the relevant numerical values and associated units using **regular expression**s. This allowed us to isolate measurements such as dimensions or weights. Additionally, any abbreviated units like "cm" or "kg" were **standardized into their full forms**, such as "centimeters" or "kilograms," to ensure uniformity across the dataset. Once the relevant values and units were extracted, we cleaned the text by removing unnecessary words and maintaining only the critical data, resulting in a concise and well-formatted output. These post processing steps ensured that the final output met the required specifications for our task.

About The Model

The Qwen-2-VL-2B-Instruct model is designed for vision-language tasks, capable of interpreting images and generating relevant textual descriptions based on prompts. It **combines a vision encoder with a language model to understand image content and generate conditioned responses**. The model's ability to integrate both visual and textual data made it ideal for our product information extraction task.

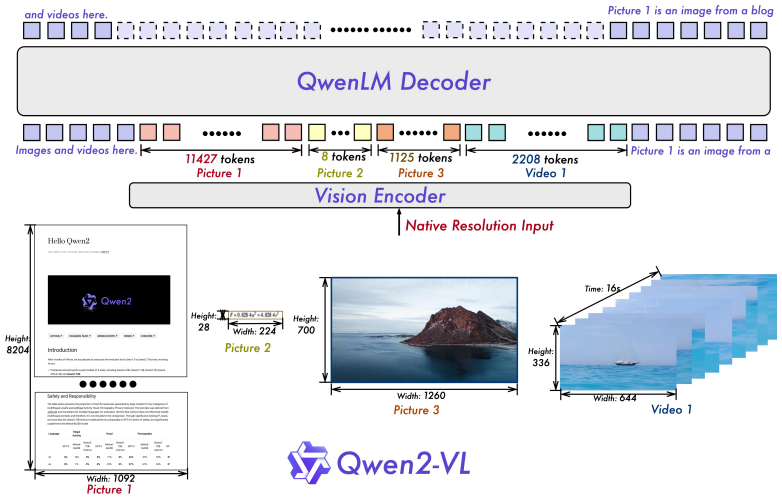

Figure 1: Qwen2-VL Architecture

Experiments

In our experiments, we used the Qwen2-VL-2B-Instruct model for product information extraction. After several rounds of optimization and postprocessing of the test data provided, we achieved an **F1 score of 0.627**. This result highlights the model's capability to perform extraction tasks with a reasonable balance between precision and recall.

Conclusion

In conclusion, the implementation of the model and the postprocessing pipeline allowed us to efficiently extract and refine relevant product information from the input images. By utilizing A6000 GPUs for faster computation and the Qwen2-VL-2B-Instruct model, we successfully generated detailed outputs for various product attributes. This approach ensured accurate extraction of product dimensions, weights, and other specifications, ultimately enhancing the usability and consistency of the data for further analysis or integration into other systems.

References

[1] 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.

[2] Qwen team. Qwen2-vl. 2024.