Methodology: Vision Feature Extraction from Images

Team Licht den Code
Siddhant, Arjun, Hetansh, Vedica

Abstract

This document presents a comprehensive methodology for extracting specific entity values from images using advanced machine learning techniques. Developed by Team Licht den Code, we detail the mathematical foundations, image processing algorithms, and vision-language model architecture used in our approach, with a focus on the Qwen2-VL model and associated utilities. We also introduce an innovative multi-modal position encoding technique, M-RoPE, to enhance the model's ability to process mixed text and image inputs.

1 Introduction

The expansion of digital marketplaces necessitates accurate and detailed product information extraction directly from images. This methodology, developed by Team Licht den Code, outlines the creation of an AI-powered system capable of identifying and extracting specific entity values such as weight, volume, dimensions, and other critical product information from images. We also introduce a novel position encoding technique for multi-modal inputs, enhancing the model's ability to process mixed text and image data.

2 Problem Formulation

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ and a set of target entities $E = \{e_1, e_2, ..., e_n\}$, our goal is to find a function $f : I \times E \to V$, where $V = \{v_1, v_2, ..., v_n\}$ represents the corresponding entity values. Each v_i consists of a numerical value and an associated unit (where applicable).

3 Methodology

3.1 Image Preprocessing

3.1.1 Smart Resizing

We employ a smart resizing algorithm to ensure optimal image dimensions while preserving aspect ratio:

$$(h', w') = \arg\min_{(h,w)} |hw - \alpha HW| \quad \text{subject to} \quad \frac{h}{H} = \frac{w}{W}, \quad h, w \in k\mathbb{Z}^+ \tag{1}$$

where H and W are original height and width, h' and w' are new dimensions, k is the dimension factor (typically 28), and α is a scaling factor to ensure the total number of pixels is within a specified range.

3.1.2 Aspect Ratio Constraint

To prevent extreme aspect ratios, we enforce:

$$\max\left(\frac{H}{W}, \frac{W}{H}\right) \leq R_{max} \tag{2}$$

where R_{max} is the maximum allowed aspect ratio (typically 200).

3.2 Vision-Language Model Architecture

We utilize the Qwen2-VL-7B-Instruct model, a large-scale vision-language model based on the transformer architecture. The model processes both image and text inputs to generate relevant textual outputs.

3.2.1 Image Encoding

The image I is encoded into a sequence of image tokens $T_I = \{t_1, t_2, ..., t_m\}$ using a vision transformer:

$$T_I = \text{ViT}(I) \tag{3}$$

3.2.2 Text Encoding

The prompt P for each entity e_i is tokenized into a sequence of text tokens $T_P = \{p_1, p_2, ..., p_l\}$:

$$T_P = \text{Tokenize}(P(e_i)) \tag{4}$$

3.2.3 Multi-modal Rotary Position Encoding (M-RoPE)

To enhance the model's ability to process mixed text and image inputs, we introduce M-RoPE, an extension of the Rotary Position Encoding (RoPE) technique. M-RoPE allows for a unified approach to position encoding across different modalities.

The core idea of RoPE is based on the rotation matrix:

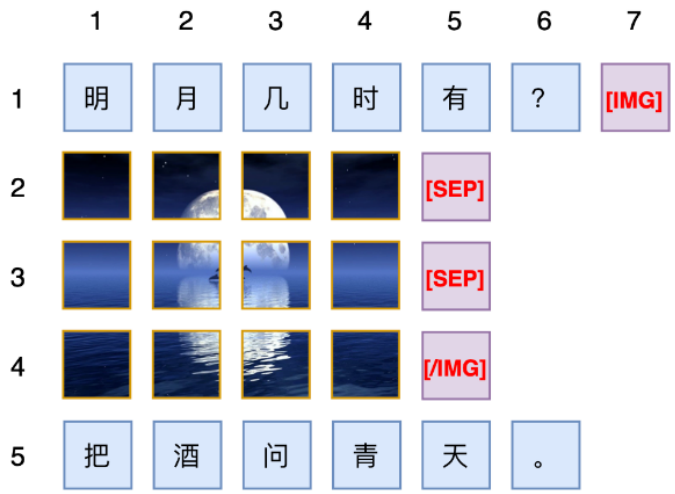
$$R_n = \begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix} \tag{5}$$

This rotation matrix has a crucial property:

$$R_m^T R_n = R_{n-m} \tag{6}$$

This property is key to RoPE's effectiveness. When applied to query-key dot products in attention mechanisms, it allows the model to capture relative positions while using absolute position encodings. This is because:

$$(R_m q)^T (R_n k) = q^T R_m^T R_n k = q^T R_{n-m} k \tag{7}$$



The Diagram displays an unified construction of 2D Position Coordinates

For multi-modal inputs, we extend this to a 2D rotation matrix:

$$R_{x,y} = \begin{pmatrix} R_x & 0 \\ 0 & R_y \end{pmatrix} = \begin{pmatrix} \cos x\theta & -\sin x\theta & 0 & 0 \\ \sin x\theta & \cos x\theta & 0 & 0 \\ 0 & 0 & \cos y\theta & -\sin y\theta \\ 0 & 0 & \sin y\theta & \cos y\theta \end{pmatrix} \tag{8}$$

This 2D extension allows us to encode both dimensions of an image simultaneously, while still maintaining the beneficial properties of the original RoPE.

To ensure compatibility with text-only inputs and allow for graceful degradation to RoPE-1D, we introduce scaling factors s and t:

$$s = \frac{wh+1}{h+1}, \quad t = \frac{wh+1}{w+1} \tag{9}$$

where w and h are the width and height of the image patch grid, respectively.

The ingenuity of this approach lies in how it unifies text and image position encoding:

1. For text, positions take the form (n, n), effectively reducing to the 1D case when no images are present.

2. For image patches, positions are scaled as (sx, ty), preserving the 2D structure of the image.

This unified approach allows the model to seamlessly handle mixed text and image inputs, while maintaining the ability to process text-only inputs efficiently. The scaling factors ensure that the transition between text and image positions is smooth, avoiding abrupt jumps in the positional space.

Furthermore, this method preserves translation invariance, a key property in many vision tasks. If we shift all positions by a constant (c, c), the relative positions remain unchanged:

$$(n+c, n+c) - (m+c, m+c) = (n-m, n-m) \tag{10}$$

This property allows the model to focus on relative positions rather than absolute positions, which is often more relevant for understanding the content and context of images and text.

3.2.4 Cross-Modal Attention

The model uses cross-modal attention to fuse image and text information:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K, and V are query, key, and value matrices derived from the combined image and text token sequences.

3.3 Inference Pipeline

3.3.1 Prompt Generation

For each entity e_i, we generate a prompt:

Extracting Information from Images

You are an AI assistant specialized in analyzing images. Your task is to extract specific information from the given image.
Please follow these instructions:
1. Locate the entity named "{entity_name}" in the image.
2. If the entity is found, extract the value associated with it.
3. The value may be followed by one of these units: {entity_units}.
4. Return only the numerical value and the associated unit, if applicable.
5. If the entity is not found, respond with "Not found".
6. For entities such as height, width, and depth, pay attention to the spatial location of the line segment and the numbers with respect to the entity.
Examples:
- For "item_weight" with "Net Wt: 500g", return "500 gram"
- For "item_volume" with "1 Cup a day", return "1 cup"
- For "height" with "Height: 10cm", return "10 centimetre"
- For "width" with "Width: 5in", return "5 inch"
- For "depth" with "Depth: 42cm", return "42 centimetre"
Guidelines:
- Be precise and only return the requested information.
- Avoid including any additional text or explanations in your response.

3.3.2 Entity Value Extraction

The model generates a sequence of output tokens $O = \{o_1, o_2, ..., o_k\}$:

$$O = \arg\max_O P(O|T_I, T_P) \tag{12}$$

3.3.3 Post-processing

We apply regex patterns to extract numerical values and units from the generated text:

$$v_i = \text{Regex}(O, \text{pattern}_{e_i}) \tag{13}$$

Regular Expression for Unit Extraction

```
(+(?:.+)?)(?:\s(?:-|\sto \s)\s+(?:.+)?)?\s*
(?:(cm|centimetre|centimeter|centimeters|centimetres)|
(mm|millimetre|millimeter|millimeters|millimetres)|
(m|metre|meter|meters)|
(\"|in|inch|inches)|
('|ft|foot|feet)|
(yd|yard|yards)|
(mg|milligram|milligrams)|
(g|gram|grams)|
(kg|kilogram|kilograms)|
(μg|mcg|microgram|micrograms)|
(oz|ounce|ounces)|
(lb|lbs|pound|pounds)|
(ton|tons|tonne|tonnes)|
(ml|millilitre|milliliter|milliliters|millilitres)|
(l|liter|litre|liters|litres)|
(cl|centilitre|centiliter|centiliters|centilitres)|
(dl|decilitre|deciliter|deciliters|decilitres)|
(μl|microlitre|microliter|microliters|microlitres)|
(gal|gallon|gallons)|
(imperial gallon|imperial gallons)|
(qt|quart|quarts)|
(pt|pint|pints)|
(cup|cups)|
(fl oz|fluid ounce|fluid ounces)|
(cu ft|cubic foot|cubic feet)|
(cu in|cubic inch|cubic inches)|
(v|volt|volts)|
(mv|millivolt|millivolts)|
(kv|kilovolt|kilovolts)|
(w|watt|watts)|
(kw|kilowatt|kilowatts)|
(a|amp|ampere|amperes))\b
```

4 Future Scope: Video Processing

For future extensions to video inputs, we propose the following methodology:

4.1 Frame Extraction

We sample frames at a specified FPS or total number of frames:

$$F = \{f_t | t = \text{round}(i \cdot \frac{T}{N}), i = 0, 1, ..., N-1\} \tag{14}$$

where F is the set of extracted frames, T is the total number of frames in the video, and N is the desired number of frames.

4.2 Frame Resizing

We apply smart resizing to each frame, ensuring consistent dimensions across the video:

$$(h'_v, w'_v) = \arg\min_{(h,w)} |hwN - \beta HWT| \quad \text{subject to} \quad \frac{h}{H} = \frac{w}{W}, \quad h, w \in k\mathbb{Z}^+ \tag{15}$$

where β is a scaling factor for video, and T is the number of frames.

5 Evaluation Metrics

We use the following metrics to evaluate our model:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively.

6 Conclusion

Our methodology combines advanced image processing techniques with a state-of-the-art vision-language model, enhanced by the novel M-RoPE position encoding. This approach allows for accurate extraction of entity values from images while maintaining flexibility for future extensions to video inputs. The introduction of M-RoPE enables seamless processing of mixed text and image inputs, potentially improving the model's performance on multi-modal tasks.

The smart resizing and aspect ratio constraints ensure optimal input for the vision-language model, while the M-RoPE technique provides a unified approach to position encoding for both text and image modalities. This integration of techniques allows for efficient and accurate extraction of entity values from product images, addressing the needs of modern digital marketplaces.

Future work will focus on extending this methodology to video inputs and further refining the M-RoPE technique for improved performance across a wider range of multi-modal tasks.