AMAZON ML CHALLENGE 2024

Team mlzon

Parth Pawar Harsh Jha Raja Yashwanth Aditya Rai

Machine Learning Approach:

The primary objective of this solution was to extract numeric values along with their corresponding units from images. We employed **PaddleOCR** for Optical Character Recognition (OCR) to extract the raw text from the images. After extracting text, the solution leverages **regular expressions** to identify numeric values and the associated units, which are then mapped to standard entities like weight, height, and volume.

Models Used:

- **PaddleOCR**:
  - A state-of-the-art OCR tool designed for high-accuracy text extraction from images.
  - It supports angle classification to detect rotated text.
  - For this project, we used the **English language model** provided by PaddleOCR to extract text from product images.
- **Regular Expressions** (Regex-based post-processing):

Once the text is extracted, we applied regex patterns to extract numeric terms followed by unit abbreviations. The specific pattern we used is:

```
(\d+\.?\d*\s?\w+)
```

  - This expression captures numbers (both integers and decimals) followed by a word (which can be a unit). This allows us to extract terms like `5 kg`, `2.5 cm`, or `1000 g`.

  - After extraction, we further processed the text using another regex pattern to extract and normalize unit abbreviations:

```
(\d+(\.\d+)?)(\s*|\b)(grams|gm|G|g|GSM|kg|Kg|KGS|cms|ug|mcg|MCG|mg|µg|oz|OZ|Lb|t|T|ml|mL|ML|mm|l|cm|CMS|kV|mV|V|W|kW)
```

- This pattern matches a number followed by a recognized unit abbreviation (e.g., `kg`, `cm`, `oz`), allowing us to replace these abbreviations with standardized units (e.g., `kilogram`, `centimetre`, `ounce`).

Experiments:

- **OCR Extraction**: We used PaddleOCR to extract text from a set of product images. The extracted text was often noisy, requiring further post-processing to isolate the relevant information.
- **Regular Expression-based Post-processing**: Using the first regex pattern, we extracted numeric terms followed by units. We then mapped these units to their respective full forms (e.g., `kg` to `kilogram`, `cm` to `centimetre`). The second regex pattern helped in normalizing the detected units.
- **Unit Standardization**: We created a unit map for each entity (such as weight, volume, etc.), ensuring that only valid units for that entity type were accepted.

Conclusion:

By combining PaddleOCR with regular expressions, we successfully extracted and standardized numeric values and their units from product images. The regex approach ensured that even abbreviated units were properly expanded and matched with their corresponding entity types. Challenges like noisy OCR outputs were handled effectively using the regex filtering approach.

Future enhancements could involve improving OCR accuracy through domain-specific training or expanding the regex patterns to cover more complex unit variations.