AMAZON ML CHALLENGE 20224

PROBLEM STATEMENT: The Goal Is To Create A Machine Learning Model That Extracts Entity Values From Images. This Capability Is Crucial In Fields Like Healthcare, E-Commerce, And Content Moderation, Where Precise Product Information Is Vital. As Digital Marketplaces Expand, Many Products Lack Detailed Textual Descriptions, Making It Essential To Obtain Key Details Directly From Images. These Images Provide Important Information Such As Weight, Volume, Voltage, Wattage, Dimensions, And Many More, Which Are Critical For Digital Stores.

ABOUT DATASET:

1. **index:** An unique identifier (ID) for the data sample
2. **image_link**: Public URL where the product image is available for download. Example link - https://m.media-amazon.com/images/I/71XfHPR36-L.jpg
   To download images use `download_images` function from `src/utils.py`. See sample code in `src/test.ipynb`.
3. **group_id**: Category code of the product
4. **entity_name**: Product entity name. For eg: “item_weight”
5. **entity_value**: Product entity value. For eg: “34 gram”

Processing Workflow

- **Text Extraction Using PaddleOCR and openbmb/MiniCPM-V-2_6:**
  We chose PaddleOCR for text extraction due to its superior prediction accuracy and faster inference times compared to other OCR techniques from the diverse images in the dataset. PaddleOCR is proficient in handling various languages and complex layouts.
  Its efficient processing of large volumes of images allowed us to promptly extract text, which was crucial given the dataset's size. Images that were captured incorrectly from paddleOCR were passed to openbmb/MiniCPM-V-2_6 to extract the relevant units from the image. Usage of the LLM would increase the inference time, so we only considered it using for incorrect paddleOCR texts from the image.

- **Refinement with Advanced Regex Techniques:**
  Following initial text extraction, we utilized advanced regular expressions (regex) to further process and refine the OCR output. Regex enabled us to:
  Identify and Extract Specific Values: We could isolate essential data such as height, width, depth, weight, weight recommendation, volume,voltage and wattage measurements types by creating regex patterns tailored to match specific keywords and numeric values.
  Clean and Format Data: Regex also helped in removing unwanted characters and formatting inconsistencies in the extracted text, resulting in a more accurate and readable dataset.
  To summerise this technique helped us to identify, extract, clean, and format specific values from the OCR output.

- **Enhanced Extraction Strategies:**
  Positional Extraction: We employed a positional extraction strategy to accurately capture depth values. This involved focusing on values based on their position in the text, such as extracting the third value in a sequence or values located near known keywords (e.g., "depth," "measurement", “net weight” etc).
  Keyword Detection: Integrating keyword detection techniques enhanced the specificity of our extraction process, allowing us to filter out irrelevant data and concentrate on values related to depth and other critical measurements.

- **Unit Conversion and Data Integration:**
  Unit Conversion: After extracting the relevant values, we converted them into SI units to standardize measurements across the dataset.
  Data Integration: The extracted and converted data were then integrated into a structured format, facilitating analysis and utilization for subsequent tasks, such as vehicle movement analysis and categorization.

- **Iterative Refinement and Optimization:**
  Throughout the process, we continuously tested and refined our methods to address challenges and improve accuracy. This iterative approach involved adjusting regex patterns, fine-tuning extraction strategies, and evaluating results to ensure optimal performance.

We considered developing different models to process various entities separately. This approach would allow each model to specialize in extracting and handling specific types of data, such as depth, height, or vehicle type. By doing so, we aimed to improve the accuracy and efficiency of entity extraction, tailoring the models to focus on the unique characteristics of each entity.

OUR WORK:

Images to Text: Images to Text.ipynb

Entity Name	**Related Work**
Height	test_height.ipynb
Width	test_width.ipynb
Depth	test_depth.ipynb
Weight	Another copy of ll_weight.ipynb

Weight Recommendation	test_rec_weight.ipynb || /www.kaggle.com/code/nsudharshanreddy/llm-maxw /edit/run/196832823
Voltage and Wattage	test_voltage.ipynb
Volume	e test_volume.ipynb

download_all.ipynb
combining all data.ipynb
split_data_for_test.ipynb