Amazon ML Challenge - Team DBkaScam

Arnav Goel, Medha Hira, Mihir Agarwal, A. S. Poornash

September 2024

1 Problem Statement

Given the images and entity name, we were tasked to create a machine-learning model that extracts entity values from images. The task was particularly challenging as we had to ensure the correct unit and value predictions.

2 Approaches Explored

We were provided with a train and test datasets of images extracted from the Amazon Database.

1. Although our initial intuition led us to approach this problem by performing a classification task for the units (based on the entity name) and a regression task for calculating the values, we quickly realized this approach would yield suboptimal results on the F1 metric, which considers True Positives, False Positives, and False Negatives. Given the maximum allowed deviation of $\pm 2\%$, this method would result in poor performance using regression.

2. We also experimented with Tesseract for OCR to extract text from images. The results were sub-optimal on the train set. Tesseract struggled with accurately parsing images, particularly when the text was too small. Additionally, it processed text strictly from left to right, often overlooking closely situated units and values, which led to numerous inaccuracies in the extracted data.

3. Finally, we explored the use of Vision Language models. This has been further elaborated in Section:3.

On taking a closer look at the dataset and performing the necessary EDA, we realized the problem at hand involved reading the dimensions of : width, depth, height, item weight, maximum weight recommendation, voltage, wattage, and item volume from the given datasets to make accurate predictions.

3 VLM

Vision Language Models (VLMs) have recently emerged as powerful tools in Optical Character Recognition (OCR), improving the accuracy and versatility of extracting text from images. These models leverage both visual and linguistic information, making them superior to traditional OCR systems, which rely solely on visual features.

Below are the VLMs that we tried for the task at hand:

- Qwen2-VL
- InternVL
- MiniCPM

Different approaches tried with VLMs:

- Zero Shot Prompting: We leveraged the model's pre-trained knowledge, focusing on crafting better prompts to achieve consistent and accurate outputs.
- Supervised Fine Tuning: We fine-tuned Qwen2-VL on a subset of the training data, but saw minimal improvement in the F1 score. This was likely due to the absence of edge cases in the data and noise, such as multiple object values in images.
- Dynamic Few Shot Prompting: We then shifted to this approach, designing **four curated examples of edge cases for each entity name** and using them as prompts. This strategy improved the model's clarity and accuracy.

During zero-shot inference with MiniCPM, we encountered noise in unit representation, with variations like "cm," "centimeter," and "centimetre," resulting in over 1,000 unique units. To address this, we used regex and a rule-based system for post-processing to standardize the format of units. Further, there were cases when the model output 2 units, one next to the value was given precedence here. We also handled edge cases, such as missing values or garbage output, and ensured the upper bound was stored for range values. We also needed additional post-processing for values that included fractions and mixed fractions. These were converted into decimal format for consistency.

Focus on prompt design helped us decrease the unique units to around 150!

Improved Prompt

Please extract the item weight and its unit of measurement from the image, providing them separately. Ensure that the unit is one of the following: {str_units}. Format your response as follows:
Value: <only the numerical value>
Unit: <unit of measurement from the specified list>

BEST PERFORMANCE: Dyanmic few shot with Improved prompt!