Amazon ML Challenge

**Team Name :** Watermelon Sugar

**Team Members :** Nikunj Jain , Manish Jain, Kushal Kuchibotla, Swetha Sridhar

**Problem Statement:** Feature Extraction from Images

**1. Introduction**

In this hackathon, the task is to develop a machine learning model that extracts key entity values from product images. The capability to retrieve structured information such as weight, dimensions, and voltage from images is essential in fields like e-commerce and healthcare, where product information is critical but often missing. This challenge requires the identification of these values directly from images through machine learning techniques, specifically optical character recognition (OCR). The extracted data must follow a specified format, and evaluation is based on the F1 score for prediction accuracy.

**2. Dataset**

The dataset provided consists of the following key columns:

* **index**: Unique identifier for each data sample.
* **image\_link**: URL where the product image can be downloaded.
* **group\_id**: Category code of the product.
* **entity\_name**: The type of product entity (e.g., item\_weight, item\_voltage).
* **entity\_value**: The target value to be extracted from images (e.g., “34 grams” for item\_weight).

**3. Methodology**

The methodology adopted for solving this problem involved a multi-step process:

* **Data Collection and Preparation**: We used the URLs in the dataset to download product images, storing them locally.
* **Optical Character Recognition (OCR)**: For extracting text from the product images, we employed **PaddleOCR**, a robust open-source model known for its accuracy in text extraction. The images contained text describing measurements like weight, dimensions, voltage, and wattage, which PaddleOCR could efficiently retrieve.
* **Pattern Matching using Regular Expressions (Regex)**: Since the extracted text is often unstructured, we used regular expressions to identify and extract numeric values corresponding to the entity (e.g., “34 grams” for weight or “12.5 cm” for dimensions). Regex is particularly useful for this task because it allows us to search for specific patterns related to various measurement units.
* **Post-Processing and Unit Standardization**: After extraction, values were standardized into allowed units, ensuring consistency in formatting (e.g., grams, centimeters, volts). For cases where multiple values were identified, the highest or most relevant value was chosen.
* **Prediction Output**: The final predictions were generated in the required format and saved in a CSV file. Each row contained an index and the predicted value in the format “x unit” (e.g., “12.5 centimetre”).

**4. Technology Stack**

We used the following tools and techniques to solve the problem:

* **PaddleOCR**: Used to perform optical character recognition on the images. It was chosen due to its high accuracy in extracting text, even from complex or distorted images.
* **Regular Expressions (Regex)**: Applied to structure the OCR output and identify specific entity values (like weights or dimensions) from the unstructured text.
* **Pandas**: For managing data input/output in CSV format, as well as handling operations related to data manipulation, such as concatenating extracted values into the final prediction format.
* **Google Colab**: The environment for running the entire pipeline, including downloading images, running OCR, and generating predictions.

**5. Experiments**

We conducted several experiments to optimize the accuracy of text extraction and prediction:

* **OCR Tuning**: Initially, we tested different OCR models, but PaddleOCR provided the most accurate results in terms of recognizing both numeric and textual information from the images.
* **Regex Optimization**: Various regular expression patterns were tried to accurately capture values and associated units (e.g., grams, volts). We experimented with multiple patterns for cases where different formats of units appeared (e.g., “cm” vs. “centimetre”).
* **Post-Processing**: After extracting the text, we experimented with different strategies for choosing the most relevant value. For example, in some cases, products displayed multiple measurements, and we developed rules to select the highest value or the one most likely to be accurate for the given entity.

**6. Results**

After implementing OCR and regex-based pattern matching, the final model was able to successfully extract entity values from images with a high degree of accuracy. Predictions were outputted in the required format, and a CSV file was generated containing the predicted values for the test dataset. The results were evaluated using the F1 score, and we achieved promising accuracy in terms of both precision and recall.

**7. Conclusion**

Our approach, combining PaddleOCR for text extraction and regular expressions for structured value retrieval, proved effective in tackling the problem of feature extraction from images. By carefully preprocessing the data, optimizing OCR performance, and refining our pattern matching techniques, we were able to meet the challenge's requirements.