**Machine Learning Approach to Extracting Entity Values from Images**

**1. Introduction**

In this project, we aimed to extract key entity values from product images. This is a common task in e-commerce, where product details like weight, dimensions, voltage, and wattage need to be extracted from images. Given that many products lack proper textual descriptions, using machine learning (ML) to identify and extract these values becomes crucial for tasks like inventory management, quality assurance, and listing accuracy. Optical Character Recognition (OCR) is used to interpret the text from images, and machine learning techniques are applied to process this text and extract relevant information.

**2. Machine Learning Approach**

The overall process involves the following steps:

1. **Data Preprocessing**:
   * The input dataset contains images of products, and the goal is to extract values related to product attributes.
   * Each image is accessed via a URL, and we apply OCR to extract any text present in the images.
   * Once the text is extracted, we clean it by removing unnecessary spaces and symbols to make it suitable for further processing.
2. **Pattern Recognition with Regular Expressions (Regex)**:
   * We employ regular expressions (regex) to identify patterns in the text that correspond to the target entities such as weight, voltage, and dimensions.
   * For example, for "item\_weight," we use a regex pattern like r'(?:item weight|weight)[\s:]\*([\d.]+\s?(?:lbs?|pounds?|kg|kilograms?|g|grams?))'. This identifies relevant keywords and captures the associated numeric value and unit (e.g., "50 pounds" or "22 kg").
   * A similar approach is taken for other attributes like depth, width, height, voltage, and wattage.
3. **Unit Conversion**:
   * After extracting values using regex, the next step involves standardizing the units. The raw text can contain multiple representations of the same unit (e.g., "lbs," "pounds," "kg"), so we map these to a standardized format.
   * A dictionary is created to map common abbreviations to their full form (e.g., 'kg' becomes 'kilogram', 'lbs' becomes 'pound'). This ensures consistency in the final output.

**3. Machine Learning Models Used**

In this project, we primarily used:

* **Optical Character Recognition (OCR)**:
  + We utilized Tesseract, an open-source OCR tool, to extract textual information from images. Tesseract effectively identifies and extracts characters, words, and numbers that appear in the image, which are then passed to the regex patterns.
* **Regex-Based Feature Extraction**:
  + Rather than using a machine learning model for classification or regression, we opted for a rule-based approach with regular expressions. This is because the task is primarily about recognizing structured patterns in the text (e.g., numbers followed by units), which can be efficiently handled by regex.

Although no supervised or unsupervised machine learning models were applied in this task (such as decision trees or neural networks), the approach aligns well with ML in terms of data preprocessing, feature extraction, and standardization.

**4. Experiments Conducted**

1. **Text Extraction Using Tesseract**:
   * We experimented with different image pre-processing techniques (e.g., changing contrast, brightness) to improve OCR accuracy. This is crucial since product labels often contain small or unclear text, which could lead to incorrect extraction.
2. **Regex Optimization**:
   * We iteratively refined our regex patterns to improve the accuracy of value extraction. For example, by testing against different formats of weights or dimensions, we ensured that the model could generalize well across various forms of input.
   * We also added flexibility to the patterns so that they could identify values in different units (e.g., pounds, kilograms) and handle potential OCR errors (e.g., missing characters).
3. **Handling Edge Cases**:
   * Several edge cases were addressed, such as:
     + Images that didn't contain any relevant information.
     + OCR errors that resulted in missing or incorrect characters.
     + Products with complex layouts, where text and numbers were not easily separable.

**5. Conclusion**

This project successfully demonstrated how Optical Character Recognition and pattern recognition techniques can be applied to extract product-related entity values from images. By leveraging OCR for text extraction and regex for structured data extraction, we were able to create a system capable of parsing a variety of attributes, such as weight, dimensions, and voltage, from product images.

The approach is particularly useful in situations where the dataset consists mainly of images with little to no accompanying textual descriptions. This solution can be extended to various applications, including inventory systems, automated product listing generation, and quality control systems.

Although no traditional ML models were used, the project highlights the importance of feature extraction in machine learning pipelines. A potential future improvement could involve using NLP models to process the extracted text more effectively or even employing deep learning techniques to enhance the accuracy of OCR on noisy or complex images.

Overall, the combination of Tesseract OCR and regex-based feature extraction provides a highly efficient and accurate solution for extracting structured data from unstructured image content.