Feature Extraction from Images: Solution Report

1. Introduction

This document outlines the approach and solution developed for the problem of **Feature Extraction from Images**. The goal of this hackathon was to create a machine learning (ML) model capable of extracting critical product details (e.g., weight, volume, dimensions) directly from images. This task is essential in fields such as healthcare, e-commerce, and content moderation, where accurate product descriptions are needed but often unavailable in text form. In such cases, extracting information directly from images becomes crucial to maintain the integrity of digital marketplaces.

2. Approach

Our solution focuses on extracting entity values such as weight, volume, voltage, wattage, and dimensions from product images using a combination of **OCR (Optical Character Recognition)** and **rule-based extraction** techniques.

Key steps in the approach:

- **OCR for Text Extraction**: PaddleOCR is used to extract text from images, including product labels, packaging, or embedded metadata. This ensures that all relevant information visible in the image is converted to machine-readable text.
- **Regular Expressions for Entity Extraction**: The extracted text is processed using predefined **regex patterns** to identify and extract specific numerical values (e.g., 12 cm, 1.5 kg) associated with the product entities (dimensions, weight, volume, etc.).
- **Unit Standardization**: The extracted values are then matched against a predefined **unit mapping** to ensure all measurements are presented in standardized units, making the data consistent and easy to interpret.

3. ML Models Used

PaddleOCR

- **Model Type**: Pre-built OCR system
- **Purpose**: To extract textual data from product images. PaddleOCR is a highly optimized OCR tool based on deep learning architectures such as **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)** for recognizing text.

Rule-Based Extraction

- **Model Type**: Not a traditional ML model but a structured pattern-matching approach using **regular expressions** to extract specific entity values from the text.
- **Purpose**: Extract numerical values (e.g., dimensions, weights) along with their units from the OCR-extracted text.

4. Experiments

The experiments primarily involved validating the performance of the solution across various product images, focusing on:

- **Text Extraction**:Verifying the accuracy of different OCR models such as **EasyOCR**, **Tesseract**, **PaddleOCR**, **Keras-OCR**, and **CRAFT** in extracting relevant text from images across different product categories.

- **Entity Extraction**: Ensuring that the regular expression patterns correctly identify and extract values corresponding to specific entities like width, height, weight, and volume.

Experimental Steps:

1. **Dataset**: A subset of product images from a larger dataset was used to evaluate the system's accuracy.
2. **Text Extraction Validation**: Manually inspect the OCR-extracted text to ensure that the extracted text correctly represents the information visible in the images.

Observations:

- **OCR Performance**: PaddleOCR performed well on clean images with clear text.
- **Entity Extraction**: The regular expression approach was effective for common units like cm, kg, and ml. However, some edge cases required additional refinement, such as handling ambiguous unit abbreviations or incorrect matches.

5. Conclusion

The implemented solution successfully extracts key product entity values from images using a combination of **OCR for text extraction** and **regular expression(regex)-based techniques for entity identification**. The use of PaddleOCR provides a solid base for text extraction, while the rule-based approach allows for flexible handling of various measurement units.

Key Takeaways:

- **Strengths**:
  - High accuracy in text extraction for clear images.
  - Rule-based extraction of entity values ensures standardized and accurate data output.
- **Limitations**:
  - The solution relies heavily on the quality of the images. Low-quality or noisy images lead to poor OCR performance.
  - A rule-based approach may not handle complex or ambiguous cases as well as a machine learning-based model designed specifically for entity extraction.