**Machine Learning Approach to Recognizing Vital Information from Images**

**Introduction**

In various industrial and scientific fields, extracting key information such as dimensions, weights, and voltage values from documents and images is essential for automating data processing tasks. This document outlines an ML-based solution to recognize such critical information from images using EasyOCR for Optical Character Recognition (OCR), combined with preprocessing and postprocessing techniques to refine the extracted data.

**Workflow Overview**

The solution is structured into distinct stages: **Loader**, **Preprocessing**, **OCR Model**, **Postprocessing**, and **Extraction**. Below is the step-by-step process, as visualized in the attached flowchart:

1. **Loader**: The first stage involves loading image file paths from a CSV dataset (train.csv) into a Pandas DataFrame. This structured format allows for easy manipulation and analysis of the dataset.
2. **Preprocessing**: Images are preprocessed before being fed into the OCR model. Grayscaling is applied to reduce the image to its intensity values, improving the performance of the OCR model. Grayscale images enhance text readability, especially for documents with varying lighting conditions.
3. **OCR Model**: EasyOCR, an open-source OCR tool, is used to recognize the text from the preprocessed images. EasyOCR supports multiple languages and fonts, making it an excellent fit for extracting textual data, such as numerical values and measurement units.
4. **Postprocessing**: Once the OCR model extracts the text, postprocessing is performed to clean the data. This includes:
   * **Lowercasing**: Converting all text to lowercase to ensure consistency.
   * **Removing Spaces and Symbols**: Regex-based operations are used to remove unnecessary spaces and symbols that may have been captured during OCR. This step is critical to prepare the text for accurate value extraction.
5. **Extraction**: The final stage focuses on extracting the relevant entity values (dimensions, weights, voltage) from the cleaned text and matching the units using a custom Python dictionary (units.py). The dictionary contains mappings for various units of measurement such as:
   * **Width**: Conversions between cm, m, mm, in, ft, and yd.
   * **Voltage**: Conversions between kV, mV, and V.
   * **Wattage**: Conversions between kW, W, etc.

The extracted and standardized values are then written to an output CSV file (output.csv).

**ML Models and Techniques**

The primary ML model used in this approach is EasyOCR for text recognition. EasyOCR uses a combination of deep learning models to detect and recognize text from images. Specifically, it employs a Convolutional Neural Network (CNN) to localize the text in the image and a Long Short-Term Memory (LSTM) network to sequentially decode the text.

The additional preprocessing and postprocessing steps rely on traditional image processing techniques (such as grayscaling) and regular expressions for text cleaning, which enhance the robustness and accuracy of the solution.

**Experiments**

**Experiment 1: Baseline with Raw Images**

In the first experiment, raw images without any preprocessing were fed into EasyOCR. The OCR model struggled with background noise and text clarity, resulting in a high error rate.

**Experiment 2: Grayscaling**

After applying grayscaling in the preprocessing step, the OCR accuracy significantly improved. Grayscaling removed unnecessary color information, allowing the OCR model to focus on text patterns.

**Experiment 3: Postprocessing with Regex**

Adding postprocessing techniques like lowercasing, removing spaces, and symbol elimination further enhanced the quality of the extracted text. This step was particularly useful for improving the accuracy of numerical values and unit recognition.

**Experiment 4: Units Matching and Replacement**

To address the variation in units across different images, a custom dictionary for units matching and replacement was developed. This allowed the solution to standardize the extracted values (e.g., converting ft to cm). As a result, the final output was consistent and ready for analysis.

**Conclusion**

The ML approach using EasyOCR, along with effective preprocessing and postprocessing techniques, proved to be a reliable method for recognizing critical information such as dimensions, weights, and voltage from images. The experiments demonstrated that preprocessing the images and cleaning the extracted text were essential steps for improving OCR accuracy. Additionally, the custom units-matching functionality ensured that the final output was standardized and suitable for further analysis. 