Machine Learning Approach for OCR-Based Entity Extraction

The approach combines Optical Character Recognition (O.C.R.) with machine learning techniques to extract and identify entities from images. It consists of three main steps: text extraction using O.C.R., text preprocessing, and the application of machine learning models to predict the entities within the extracted text.

1. O.C.R. Extraction

Optical Character Recognition (O.C.R.) is the foundational step in this workflow. It involves using a neural network-based method to detect and recognize text within images. In this project, the EasyOCR library was utilized for its efficiency and ability to handle various fonts, sizes, and orientations of text within images.

- *EasyOCR*: A deep learning-based O.C.R. tool that provides high accuracy for text recognition. It is capable of recognizing text in multiple languages and handling complex layouts.

- *Process:* EasyOCR processes each image in the dataset, extracts textual content, and converts it into a machine-readable format. This step transforms visual data into a textual format, which can then be analyzed using machine learning techniques.

- *Output*: The extracted text from each image is saved in a structured format (CSV file). This text serves as the input for subsequent preprocessing and model training steps.

2. Text Preprocessing

After extracting text from images, the next step is preprocessing to prepare the text for the machine learning model. Text preprocessing is crucial to ensure that the model receives clean and standardized input.

- *Normalization*: Converts the extracted text to lowercase and removes special characters. This helps in reducing variability in the text data, making it more uniform for model training.

- *Numerical Value Extraction*: Many entities, such as measurements and quantities, contain numerical values. Regular expressions are employed to extract numbers and their associated units from the text. This is particularly useful for predicting entities that are represented by numerical values, such as "50 kg" or "120 mL".

- *Output*: The preprocessed text is a standardized and normalized version of the original O.C.R. output, focusing on the relevant information needed for entity prediction.

3. Machine Learning Model

With the preprocessed text data in hand, a machine-learning model is applied to predict entities. The machine learning model used in this approach is a Naive Bayes classifier integrated within a TF-IDF pipeline.

- *TF-IDF Vectorization*: Converts the preprocessed text into numerical feature vectors. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of words in a document relative to a collection of documents. It helps capture the relevance of words in the context of the entity prediction task.

- *Naive Bayes Classifier*:

  - A probabilistic classifier known for its efficiency and performance in text classification problems.

  - It works on the principle of Bayes' theorem and is particularly suited for tasks where the features (in this case, words or terms in the text) are independent.

  - In this workflow, `MultinomialNB` is used, which is designed for multinomially distributed data, typically the case for word counts in text classification.

- Training Process:

  - The OCR-extracted text from the training dataset is used to fit the Naive Bayes model. Labels (entity types) associated with the training data guide the model in learning patterns within the text that correspond to different entities.

- Prediction:

  - The model is applied to the test dataset to predict entities once trained. The model identifies the most likely entity class for each text piece based on learned patterns.

Advantages of this Approach

- Automation: This pipeline automates the process of extracting and predicting entities from images, eliminating the need for manual data entry.

- Scalability: The combination of O.C.R. and machine learning allows the system to process a large number of images quickly and accurately.

- Flexibility: EasyOCR can handle various types of images with different text layouts, and the machine learning model can be retrained to adapt to new entity types.

In summary, this approach leverages the strengths of O.C.R. for text extraction and machine learning for entity recognition. Transforming visual data into structured information provides an efficient solution for extracting entities from images in various applications.