# **Feature Extraction from Images: ML Approach and Experiments**

## **1. Introduction**

In this hackathon, we aimed to develop a machine-learning model capable of extracting entity values such as weight,height,volume,etc. from images.

## **2. ML Approach**

Our approach combines Optical Character Recognition (OCR) and Natural Language Processing (NLP) techniques to extract and interpret relevant information from images. The process involves several key steps:

1. **Image Preprocessing**: Enhance image quality to improve OCR accuracy.
2. **Text Extraction**: Use OCR to extract text from the images.
3. **Entity Recognition**: Apply NLP techniques to identify and extract specific entities from the extracted text.
4. **Post-processing**: Refine and validate the extracted information.

## **3. ML Models and Technologies Used**

### **3.1 Optical Character Recognition (OCR)**

After trying and testing many alternative engines and packages to preprocess the image and extract relevant information, we came to the conclusion that easyOCR is the most effective in this case.

### **3.2 Natural Language Processing (NLP)**

For entity recognition and extraction, we employed spaCy, an industrial-strength NLP library. We used the following components:

1. **spaCy's English language model**: We used the en\_core\_web\_sm model for basic NLP tasks.
2. **Custom EntityRuler**: We created a custom entity recognition component to identify specific entities like weight, volume, voltage, and wattage.
3. **Rule-based matching**: We implemented custom rules to capture entities that might be missed by the pre-trained model.

## **4. Experiments**

We conducted several experiments to optimize our feature extraction pipeline:

### **4.1 OCR Optimization**

1. **Image Preprocessing**: We tested various preprocessing techniques including resizing, grayscale conversion, and contrast enhancement. We found that adaptive thresholding combined with noise reduction yielded the best results for our dataset.
2. **OCR Configuration**: We experimented with different confidence thresholds and allowed lists to improve the accuracy of text extraction. A confidence threshold of 0.5 and an allowed list of alphanumeric characters and common symbols provided the best balance between precision and recall.

### **4.2 Entity Extraction Refinement**

1. **Pattern Matching**: We developed and iteratively refined RegEx patterns for each entity type (weight, volume, voltage, wattage). We found that more specific patterns (e.g., r'(\d+(?:\.\d+)?)\s\*(?:kg|KG|g|G|lbs?|LBS?|ounces?|oz)' for weight) yielded better results than generic ones.
2. **Context-aware Extraction**: We implemented a context-aware approach for ambiguous cases. For example, when extracting weight, we prioritized values near keywords like "weight" or "mass".
3. **Unit Normalization**: We developed a unit conversion system to normalize all extracted values to a standard unit (e.g., converting all weights to kilograms).
4. **Case Sensitivity**: By using LOWER: in the RegEx pattern, we ensure that no values like 500mG, 200FT,etc aren’t left out.
5. **Priority Order:** In certain cases when there are multiple entity values for the same entity\_type, we employed certain priority orders to get the relevant result. Some of those cases are:
6. **Parenthesis:** For the cases when the same entity is measured in various units, we issued a priority order in spacy to ensure that the value within the parenthesis is taken of higher priority.  
   Ex. 35Oz is taken as the output in this !0g (0.35 Oz)
7. **Maximum value:** When there are multiple entities which add up to give a single entity in the image, we ensured a algorithm to ensure that the maximum entity value will be returned.  
   Ex. Ingredients of a cereal have 10g,100g,200g,50g,etc. and the weight of the cereal box is 400g. We return 400g in this case.

## **5. Results and Conclusion**

1. **Importance of Preprocessing**: Image preprocessing significantly improved OCR accuracy, especially for challenging images.
2. **Context Matters**: Incorporating contextual information and domain-specific rules greatly enhanced the accuracy of entity extraction.
3. **Hybrid Approach**: Combining rule-based methods with machine learning models (like spaCy's NER) provided a robust solution capable of handling various image and text formats.
4. **Continuous Refinement**: Regular error analysis and rule refinement were crucial for improving the system's performance over time.

## **6. Future Work**

While our current approach shows promising results, there are several avenues for future improvement:

1. **Deep Learning for Entity Recognition**: Explore deep learning models specifically trained for entity recognition in product images.
2. **Multi-modal Learning**: Incorporate both image and text features for more accurate entity extraction.
3. **Active Learning**: Implement an active learning pipeline to continuously improve the model with human feedback.
4. **Expanded Entity Types**: Extend the system to recognize and extract additional entity types relevant to e-commerce and other domains.

By continuing to refine and expand this approach, we can create increasingly accurate and versatile systems for extracting critical information from product images, thereby enhancing the capabilities of digital marketplaces and other image-intensive applications.