Team - priml

Amazon ML Challange

Aryan Jha[1], Rudransh Agarwal[1], Piyush Kumar[1], and Shubhendu Pandey[1]

[1]Indian Institute of Technology Kharagpur

September 18, 2024

1 Problem Statement

The task is to develop a machine learning model that extracts specific entity values, such as weight, volume, or dimensions, from product images. This is essential for sectors like e-commerce, where detailed product information is often missing. The dataset includes images with labels for entity names and values (e.g., "34 grams"), though the test set lacks values, which the model must predict. The output should be in a specific format (e.g., "x unit") and validated using a provided sanity checker. The model's performance will be evaluated using the F1 score, emphasizing precision and recall in extracting correct values from the images.

Keywords
computer vision, ocr, e-commerce, transformers

2 Methodology

2.1 OCR Detection

The first step in solving the feature extraction problem involved applying Optical Character Recognition (OCR) to extract text from product images. We experimented with both PaddleOCR and EasyOCR, and the results showed that PaddleOCR performed significantly better. PaddleOCR exhibited higher accuracy in recognizing complex characters, especially from low-quality or varied-resolution images, whereas EasyOCR struggled with extracting precise text, particularly for smaller fonts and complex layouts. The superior performance of PaddleOCR made it the preferred choice for further processing in this project.

A major challenge we faced in extracting dimensional features using OCR was its inability to differentiate between length, width, and height. Since OCR is designed to recognize and extract text, it treats all numerical data equally, without understanding the context or spatial relationships between the values. For example, when an image contains multiple dimensions such as "10 cm x 5 cm x 3 cm", OCR can extract the numbers but lacks the capability to label or distinguish which value corresponds to length, width, or height. This limitation arises because OCR is not built to interpret the semantic meaning or the physical arrangement of these values, making it unsuitable for tasks that require an understanding of specific dimensional attributes.

Criteria	**PaddleOCR**	**EasyOCR**
Accuracy	High	Moderate
Performance	High	Low
Speed	Slow	Moderate

Table 1: Comparison of PaddleOCR and EasyOCR

2.2 YOLO

To overcome the limitations of OCR in detecting dimensional features, we applied YOLO (You Only Look Once) for dimension detection. YOLO helped in identifying and creating bounding boxes around specific regions in the images that likely contained dimensional information. By isolating these areas, we could then feed the contents of the bounding boxes to PaddleOCR, which allowed for more accurate extraction of dimensions like length, width, and height. This two-step process of using YOLO for object detection and PaddleOCR for text recognition significantly improved the accuracy of extracting dimensional data from the images.

To process the images using YOLO, we first annotated the height and width data using CVAT (Computer Vision Annotation Tool). In this step, we manually created bounding boxes around the height and width information present in the images. Once the annotations were complete, we used this labeled data to train and

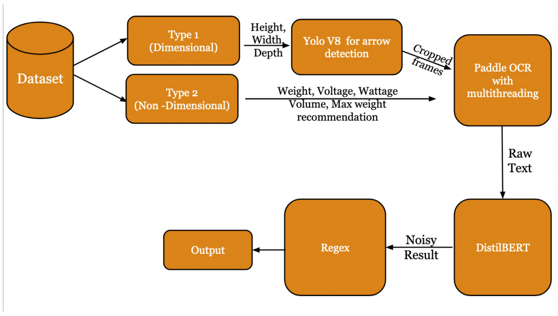
Dataset
Type 1 (Dimensional)
Type 2 (Non -Dimensional)
Height, Width, Depth
Yolo V8 for arrow detection
Cropped frames
Weight, Voltage, Wattage Volume, Max weight recommendation
Paddle OCR with multithreading
Raw Text
DistilBERT
Noisy Result
Regex
Output

Figure 1: Our Pipeline

fine-tune YOLO specifically for detecting these dimensional features. This fine-tuning enabled YOLO to accurately identify the regions containing height and width, allowing for more precise extraction of the dimensions using OCR.

2.3 Regex Matching

we also employed Regex matching to further refine the OCR output. This step involved analyzing the text extracted by OCR for common errors, such as misinterpreted characters or incorrect formatting. By applying regular expressions, we could identify and correct these typical mistakes, such as erroneous unit conversions or misplaced decimal points, ensuring that the final text data was more accurate and reliable for subsequent processing. This added layer of validation helped improve the overall precision of the extracted dimensional features.

2.4 Transformer

To enhance the accuracy of feature extraction, particularly for identifying specific dimensions like height, width, and other entity values, we incorporated a question-answer transformer model, DistilBERT. After using OCR to extract raw text from the images, we fed this text into DistilBERT, formulating it in a question-answer format. The model was prompted with specific questions aimed at identifying the dimensional features within the text, such as "What is the height?" or "What is the width?". DistilBERT, known for its efficiency and contextual understanding, was able to interpret the text and output the relevant feature based on the prompt. This method allowed us to precisely map the extracted OCR text to specific features, addressing the ambiguity that can arise from OCR outputs and ensuring a higher degree of accuracy in identifying the correct entity values from the images.

3 Discussion

To tackle the challenge of extracting dimensional features from images, we employed a multi-step approach combining several advanced techniques. Initially, we used YOLO for dimension detection, annotating height and width data with CVAT and training YOLO to create precise bounding boxes around these features. The extracted regions were then processed by PaddleOCR, which demonstrated superior text recognition capabilities compared to other OCR tools. To accurately interpret the extracted text, we integrated DistilBERT, a question-answer transformer, which was prompted to identify specific features such as height or width. Additionally, we employed Regex matching to detect and correct common OCR errors, further enhancing the accuracy of the extracted information. This comprehensive methodology ensured a robust and precise extraction of dimensional data from the images.

Conclusions

We evaluated the effectiveness of different techniques for extracting dimensional features from images, and the results varied significantly. Using only EasyOCR in combination with a transformer model achieved an F1 score of 0.29, indicating limited accuracy. Integrating YOLO for dimension detection with PaddleOCR and the transformer model improved the score to 0.56, reflecting a more accurate extraction process. The best performance was achieved by adding Regex matching to the YOLO + PaddleOCR + Transformer setup, resulting in an F1 score of 0.635. This indicates that the combination of YOLO for bounding box detection, PaddleOCR for text extraction, the transformer model for feature identification, and Regex for error correction provides the most robust solution for accurately extracting dimensional features from images.

Technique	**F1 Score**
EasyOCR and Transformer	0.29
YOLO + PaddleOCR + Transformer	0.56
YOLO + PaddleOCR + Transformer + Regex	0.635

Table 2: Comparison of F1 Scores for Different Techniques