Machine Learning Approach for Amazon 2024 ML Hackathon: Feature Extraction from Images

1. ML Approach

Primary Model: InternVL

Our main model, **InternVL** (GitHub Repository), is a multi-modal open source large language model, making it particularly well-suited for extracting entities such as dimensions, weights, and other product-specific information directly from images.

InternVL's ability to interpret visual and textual data in context allowed us to extract entity values directly from images where visual context, such as spatial relationships, was crucial.

Supplementary OCR Tools

For cases where **InternVL** could not provide sufficient predictions or the entity extraction required detailed textual recognition, we utilized a series of open-source OCR tools:

1. **EasyOCR** (GitHub Repository): Initially used to perform text extraction from images.
2. **PaddleOCR** (GitHub Repository): Employed as a backup to cross-verify the results from EasyOCR.
3. **AppleOCR** (GitHub Repository): Selected as the final step due to its superior accuracy in extracting the most amount of information in string form.

Combined Model Approach

After comparing the outputs from all three OCR tools, **AppleOCR** consistently provided the most accurate and detailed textual data. Thus, we integrated the outputs from InternVL with the results obtained from AppleOCR to develop a robust final model, ensuring maximum coverage and precision in extracting entity values from the provided images.

2. Experiments Conducted

Our experimental process was designed to validate the effectiveness of InternVL and the OCR tools for extracting entity values from images. The steps were as follows:

2.1 Exploratory Data Analysis (EDA)

We conducted an extensive **Exploratory Data Analysis (EDA)** on the dataset to identify the distribution of various entity types. The frequency analysis revealed that the majority of the

entities were related to dimensions, which informed our decision to prioritise InternVL. **InternVL** is particularly suited for dimension-related extractions due to its deep understanding of spatial and contextual relationships between visual elements and text.

2.2 Model Selection and Evaluation

- **InternVL** was utilised to extract entities directly from images, particularly focusing on those with high occurrences of dimension entities, given its specialisation in visual-language tasks.
- For entities where InternVL predictions were insufficient, we tested multiple OCR tools (EasyOCR, PaddleOCR, and AppleOCR). After careful evaluation, **AppleOCR** provided the most reliable and comprehensive results.

2.3 Integration and Fine-Tuning

- We integrated the predictions from **InternVL** with the most accurate OCR results (from AppleOCR). This approach allowed us to build a comprehensive model that leveraged the strengths of both visual-language understanding and textual data extraction.
- The combined model was fine-tuned by adjusting parameters and optimizing for specific cases where outliers or inconsistencies were detected. This ensured robustness against noisy data and partial information.
- We made multiple attempts at improving by focusing on the more important entities as per their distribution and extracting as much relevant information we could using Regular Expressions(regex).

2.4 Handling Outliers

To handle outliers, we performed **additional data preprocessing and normalisation**, which involved removing noise, correcting skewed data distributions, and refining entity extraction rules to handle rare or ambiguous cases effectively.

The above was especially true in the case of the depth entity where we had resorted to predicting no value in many cases upon inspection of multiple samples.

4. Conclusion

By combining **InternVL** with **AppleOCR**, we achieved a robust solution for extracting entity values from images in this hackathon. InternVL's ability to interpret visual-language relationships made it ideal for extracting complex entities like dimensions, while AppleOCR provided superior performance in capturing detailed textual data.

This approach underscores the importance of selecting the right models and tools for specific data characteristics and highlights the value of integrating multiple methodologies to address diverse data extraction challenges effectively.