International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

Structured Entity Extraction From Product Images Using Fine-Tuned Vision-Language Models For Digital Marketplaces

Dr. T. Jalaja[1], Dr. T. Adilakshmi[2], Vamsi Krishna Desineedi[3], Spoorthi Vadlakonda[4]

[1] Assistant Professor, Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, India, jalaja.t@staff.vce.ac.in

[2] Professor & HOD, Department of Computer Science and Engineering, Vasavi College Engineering, Hyderabad, India, t_adilakshmi@staff.vce.ac.in

[3,4] Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, India, vamsi14102003@gmail.com[3], spoorthivadlakonda8@gmail.com[4]

Abstract

In the evolving landscape of e-commerce, product listings often lack consistent and structured metadata such as weight, dimensions, or voltage – all critical for accurate cataloging and comparison. This project addresses the challenge by developing an AI-based system that can extract such entity values directly from product images. Leveraging the PaLI-Gemma vision-language model, fine-tuned using the Low-Rank Adaptation (LoRA) method, the system is trained on 5,000 annotated product images made publicly available by Amazon. The model receives both the image and a prompt specifying the desired attribute (e.g., "What is the weight?"), and returns a structured output. With entity-specific prompts and efficient fine-tuning, the system demonstrates a significant performance improvement over the base model, achieving a 0.70 F1 score on a held-out test set. This solution automates the metadata extraction process, offering a scalable and precise alternative to manual annotation in digital marketplaces.

Keywords: *Vision-Language Models, Entity Extraction, PaLI-Gemma, LoRA, E-commerce Automation*

1. INTRODUCTION

In digital marketplaces, structured product metadata—such as weight, volume, dimensions, and voltage—is crucial for catalog organization, search, and comparison. However, many listings lack this data, especially when descriptions are incomplete or must be manually curated. This project introduces an AI-based system that automatically extracts such attributes from product images using a fine-tuned vision-language model, PaLI- Gemma-3B. By framing the task as visual question answering, the model interprets an image alongside a prompt (e.g., "What is the voltage?") and returns the relevant entity in structured form. Fine-tuning was performed using Low-Rank Adaptation (LoRA) on a publicly available dataset of 5,000 annotated Amazon product images, enabling efficient and domain-specific training. The result is a scalable, accurate solution for enriching product catalogs, reducing manual effort, and improving data consistency in large-scale e-commerce systems.

2. LITERATURE REVIEW

In recent years, the use of vision-language models (VLMs) has marked a significant shift in tasks that require joint understanding of images and text. Models like PaLI and PaLI-Gemma have shown remarkable success in document understanding and visual question answering, offering strong capabilities for extracting structured information from semi-structured product images [1]. These models benefit from large-scale pre-training and are designed to interpret both visual layouts and linguistic context, making them suitable for complex e-commerce use cases.

Adapting such large models to specific downstream tasks, however, remains computationally intensive. To mitigate this, Hu et al. introduced LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning technique that injects low-rank matrices into the model's architecture without updating the bulk of its

weights [2]. This has enabled resource-constrained training while retaining performance. Building on this, NoRA (Nested LoRA) was proposed to improve flexibility in tuning across multiple tasks by introducing hierarchical adaptation layers [3].

Traditional pipelines for entity extraction typically relied on OCR engines paired with rule-based post-processing logic. While functional in many settings, OCR approaches often falter when faced with challenges like irregular layouts, low-resolution images, or skewed text, which are common in real-world product listings [4]. As a response, modern approaches increasingly frame entity extraction as a visual question-answering (VQA) problem

— where the model is prompted to extract specific attributes like “weight” or “dimensions” directly from the image content [5].

Beyond PaLI and its derivatives, other vision-language models like VisualBERT have demonstrated strong performance in multi-modal reasoning by tightly integrating visual and textual embeddings [6]. These models are often optimized using downstream benchmarks like SuperGLUE, which evaluate a model’s general understanding across multiple NLP tasks [8], showing their potential for adaptable entity-level understanding in hybrid vision- text domains.

Additionally, researchers have explored layout-aware models such as LayoutLM and Donut, which go beyond token-level representations by incorporating spatial structure into the learning process. These models are particularly useful when handling documents or product images with rich layout information, enabling more accurate extraction from tables or multi-column formats.

In parallel, foundational advancements like EfficientNet have contributed to more scalable and lightweight visual backbones, improving image encoding efficiency without sacrificing representational power — a crucial factor for real-time product attribute extraction [7].

Overall, the literature reveals a steady move from rigid, rule-based systems to flexible, prompt-adaptive architectures. By integrating pretrained VLMs with lightweight tuning methods like LoRA, and drawing from advances in vision and language understanding, recent approaches achieve high accuracy, scalability, and robustness — all essential for real-world e-commerce deployments.

3. METHODOLOGY

The goal of this work is to extract structured entity values such as item weight, volume, dimensions, voltage, and wattage directly from product images using a vision-language model. Specifically, we use the PaLI-Gemma model, which is pre-trained on large-scale image-text datasets, and fine-tune it with a parameter-efficient approach called Low-Rank Adaptation (LoRA) on a domain-specific dataset of product images. This transforms the task into a Visual Question Answering (VQA) problem, where the model is prompted to answer entity-specific questions based on the input image.

The dataset used in this project was publicly released by Amazon and contains around 5,000 annotated product images, each associated with metadata in a structured format. Each data point includes an image_link (URL to the product image), a group_id (identifier for grouping multiple images of the same product), an entity_name (the attribute being measured, such as item_weight or item_volume), and an entity_value (the ground truth value of that entity). Multiple images may exist for a single product, requiring dynamic prompting based on the target entity. For instance, the prompt for item_weight would be "What is the weight?", while for volume, it would adjust accordingly.

To fine-tune the PaLI-Gemma model for this task, LoRA is employed to reduce computational requirements while maintaining performance. Instead of updating all model parameters, LoRA introduces trainable rank decomposition matrices into specific layers. This makes the fine-tuning process lightweight and scalable, even when using large transformer-based architectures.

The training objective is to maximize the F1 score, which balances precision and recall across the predicted entity values. The evaluation is performed on a held-out test set, and the complete workflow—from data preprocessing and dynamic prompt generation to LoRA-based fine-tuning and evaluation—is illustrated

in **Figure 3.1**.

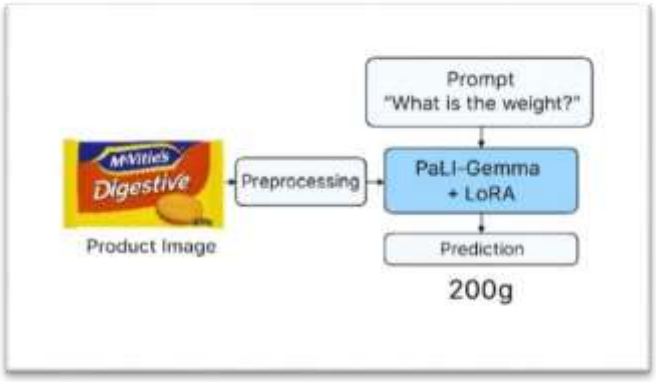

Figure 3.1. Workflow of the proposed system for structured entity extraction from product images.

4. IMPLEMENTATION

The implementation of this project involves multiple phases, including data preprocessing, model selection, prompt engineering, fine-tuning with LoRA, and finally, evaluation. Each component is crucial in adapting the PaLI-Gemma vision-language model to the task of structured entity extraction from product images.

4.1 Data Collection and Preprocessing

The dataset used in this project was sourced from a publicly available product metadata collection released by Amazon, which contains approximately 5,000 product images, each paired with structured metadata. Each data point includes an **image_link** (the URL of the product image), a **group_id** (an identifier to group different views of the same product), an **entity_name** (the target attribute to extract, such as item_weight or item_volume), and an **entity_value** (the ground truth value for that entity).

To prepare the data for model training, several preprocessing steps were performed. First, all images were downloaded and stored locally using an automated script. Only samples with valid, non-empty **entity_value** fields were retained for training. Next, the entity values were standardized to ensure consistency across units, such as converting all weight values into a single unit (e.g., grams, kilograms, ounces). Finally, each image-entity pair was transformed into a triplet consisting of the image, a prompt, and the corresponding answer. For instance, for an image of a product (e.g., a bag of flour), the prompt would be "What is the weight?" and the answer would be the corresponding ground truth value, such as "500 grams." This triplet format aligns with the Visual Question Answering (VQA) task, where the model is trained to focus on extracting the specific entity being queried.

4.2 Prompt Engineering

Since multiple entities exist for a single product, a dynamic prompt mechanism was developed to generate tailored prompts for each entity. Fixed templates were created for common entity types to ensure consistency and clarity in model input. The prompt templates used for each entity are summarized in **Table 4.2.1**.*Table 4.2.1. Prompt templates for different entity types*

Entity Name	Prompt Template
item_weight	What is the weight?
item_volume	What is the volume?
wattage	What is the wattage?
voltage	What is the voltage?

This ensures the model is consistently asked the right type of question based on the entity_name, making the task aligned with standard VQA formats.

4.3 Model Selection and Architecture

The base model used is **PaLI-Gemma**, a multilingual vision-language model designed for document and product image understanding. It combines a vision encoder with a language decoder to handle image-text tasks.

For this task, the **google/paligemma-3b-ft-docvqa-448** variant was selected as the starting checkpoint. It is pre- trained on DocVQA-like tasks, making it suitable for visual entity extraction.

However, full fine-tuning of such a large model is resource-intensive. To address this, **Low-Rank Adaptation (LoRA)** was employed.

4.4 Fine-Tuning with LoRA

Low-Rank Adaptation (LoRA) was employed to fine-tune the PaLI-Gemma model efficiently by introducing a small number of trainable parameters while keeping the majority of the original model weights frozen. This approach significantly reduces the memory and computational requirements associated with fine-tuning large transformer-based models.

LoRA was specifically applied to the language decoder layers, and only 0.5–1% of the total model parameters were updated during training. To further optimize resource usage, mixed-precision training (FP16) was used to reduce memory consumption. The training process was conducted for a few epochs, with early stopping based on the F1 score to prevent overfitting and ensure optimal model performance.

The fine-tuning was carried out using several libraries and tools: (i) **Transformers** and **PEFT** libraries from Hugging Face, (ii) **BitsAndBytes** for quantization-aware training, and (iii) **PyTorch Lightning** to manage training workflows efficiently.

The training setup involved the following parameters: (i) **Batch Size:** 8, (ii) **Optimizer:** AdamW, (iii) **Learning Rate:** 2e-5, (iv) **Scheduler:** Linear Warmup, and (v) **Loss Function:** Cross Entropy Loss (token-level).

4.5 Inference Pipeline

During inference, the model receives an input image along with a corresponding entity-based prompt, which guides the model in generating the predicted entity value in textual form. A custom wrapper around the PaLI- Gemma model is used to facilitate the inference process, which includes several key steps: (i) prompt injection,

(ii) tokenized input creation, (iii) output decoding, and (iv) post-processing (such as unit normalization).

For example, if the input is an image of a shampoo bottle and the prompt is "What is the weight?", the model's output could be "200 milliliters." This output is then normalized to a consistent unit format.

Additionally, a separate script was developed to evaluate the model across the test dataset, assessing its performance for each entity type. This evaluation ensures that the model's predictions are aligned with the expected ground truth values for each entity.

5. RESULTS

The performance of the proposed product entity extraction system was evaluated using a publicly available dataset released by Amazon. This dataset comprises 5,000 images annotated with structured attributes such as *item_weight* and *item_volume*, along with corresponding ground-truth values. We assessed the effectiveness of our approach by comparing the base model's predictions against the fine-tuned PaLI-Gemma model, which was enhanced using LoRA-based adaptation techniques.

The models were evaluated using standard metrics commonly employed for entity extraction and information retrieval tasks: F1 Score, Precision, Recall, Exact Match, and Accuracy. The results are summarized in **Table 5.1**.

Table 5.1. Performance comparison between base and fine-tuned models

Metric	Base Model	Fine-Tuned Model
F1 Score	0.5969	0.7080
Precision	0.4255	0.5480
Recall	1.0000	1.0000
Exact Match	2119 / 5000	2729 / 5000
Accuracy	0.4255	0.5480

As shown, the fine-tuned model significantly outperformed the base model across all metrics. Notably, there was an approximate 11% absolute increase in both accuracy and F1 score, indicating improved reliability and better generalization on unseen product images. Precision also improved substantially, reflecting the model's enhanced ability to generate exact entity values more consistently.

To further visualize the comparative performance, a bar chart was generated **(Figure 5.1.)**, which clearly illustrates the improvement across key evaluation metrics.

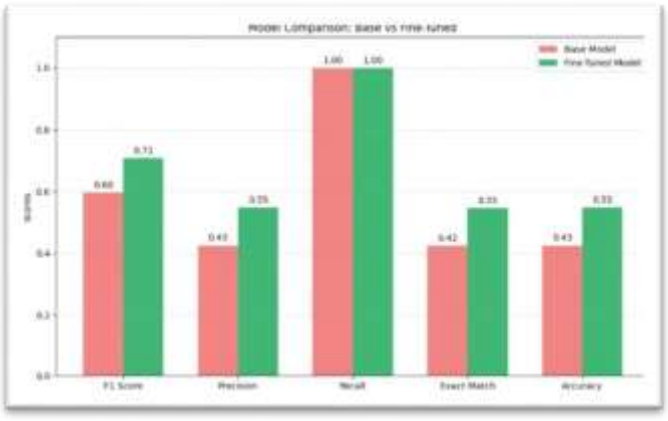

Figure 5.1. *Comparison of evaluation metrics between base model and fine-tuned model*

In addition to quantitative evaluation, qualitative insights were gathered through real-time model outputs. A few sample predictions are shown in Figures 5.2 to 5.4 to illustrate how the model extracts structured information directly from raw product images. For instance, Figure 5.2 shows the predicted entity "*1400 milligram*" extracted from the label of a Psyllium Husk supplement bottle. Figure 5.3 demonstrates the model correctly identifying "*2.75 inch*" as the diameter from an image of a metal strainer with annotated dimensions. In Figure 5.4, the model predicts "*265 volt*" from an outdoor LED lighting product by interpreting textual cues embedded in the design. These visual results highlight the model's ability to locate and extract relevant information in varied contexts, even with diverse fonts, angles, and packaging layouts. Such performance underscores its potential for real-world applications like catalog automation and inventory enrichment in digital commerce platforms.

Figure 5.2: Supplement Bottle with Predicted Weight

Figure 5.3: Strainer with Predicted Diameter

Figure 5.4: LED Light with Predicted Voltage

6. CONCLUSION AND FUTURE SCOPE

In this research, we presented a structured approach to extract product-specific entities from images using the PaLI-Gemma vision-language model, which was further enhanced through parameter-efficient fine-tuning (LoRA). By leveraging an open-source dataset released by Amazon, which contains diverse retail product images along with associated attributes, we addressed the challenge of generating entity values such as weight, volume, and size directly from visual data.

Our fine-tuned model demonstrated substantial improvements across key evaluation metrics, including F1 Score, Precision, Exact Match, and Accuracy, with particularly notable gains in prediction precision and semantic correctness. The results clearly indicate the potential of multimodal transformer architectures when adapted to real-world information extraction tasks, particularly in domains like e-commerce, where visual content is abundant but often underutilized.

The implementation successfully incorporated modular preprocessing, dynamic prompt generation, and structured output parsing. Furthermore, the evaluation framework provided reliable and interpretable performance indicators, both quantitatively and visually. These contributions serve as a solid foundation for more advanced applications involving visual understanding and structured reasoning.

Looking ahead, several promising directions exist for extending this work. First, expanding the model to handle multi-entity extraction from a single image in a single pass could significantly enhance its efficiency. Additionally, incorporating multilingual prompt templates would allow the model to generalize across different geographic markets. Integrating Optical Character Recognition (OCR) could complement visual cues by capturing embedded text features from labels or packaging. Furthermore, applying this fine-tuned pipeline to other retail domains, such as clothing, electronics, or perishables, could address the diverse

nature of product specifications. Finally, the deployment of this system within real-time platforms, such as chatbots or intelligent assistants in retail environments, could enable dynamic querying and interpretation of image content via APIs.

In conclusion, this research demonstrates the feasibility and effectiveness of adapting foundation models for domain-specific visual information extraction. With further refinements, the proposed system holds significant potential to contribute to the development of more intelligent, scalable, and automation-driven retail data infrastructures.

REFERENCES:

[1] X. Chen, B. Zoph, G. Ghiasi, V. Vasudevan et al., "PaLI: A Jointly-Scaled Multilingual Language-Image Model", Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Google Research (USA), June 19–24, 2022, pp. 12445–12456.

[2] E. J. Hu, Y. Shen, P. Wallis, H. Shen et al., "LoRA: Low-Rank Adaptation of Large Language Models", Proceedings of the 40th International Conference on Machine Learning (ICML), Microsoft Research (USA), July 18–24, 2021, pp. 12154–12165.

[3] C. Lin, L. Li, D. Li, and H. Xiong, "NoRA: Nested Low-Rank Adaptation for Efficient Fine-Tuning of Large Models", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Huawei Technologies (China), January 2024, pp. 315–324.

[4] J. Kim, S. Yun, J. Park, and H. Lee, "OCR-VQA: Visual Question Answering by Reading Text in Images", Proceedings of the International Conference on Computer Vision (ICCV), KAIST (South Korea), October 11–17, 2021, pp. 10515–10524.

[5] T. Kawano, R. Higashinaka, and H. Kameko, "Understanding Product Images with Vision-Language Models for E-Commerce", Proceedings of the 7th International Workshop on Document Analysis Systems (DAS), Rakuten Institute of Technology (Japan), February 2023, pp. 64–72.

[6] Y. Li, H. Su, D. Xu, and L. Fei-Fei, "VisualBERT: A Simple and Performant Baseline for Vision and Language", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stanford University (USA), November 16–20, 2020, pp. 2535–2544.

[7] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", Proceedings of the 36th International Conference on Machine Learning (ICML), Google AI (USA), June 9–15, 2019, pp. 6105– 6114.

[8] A. Wang, Y. Pruksachatkun, N. Nangia et al., "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems", Advances in Neural Information Processing Systems (NeurIPS), New York University (USA), December 2019, pp. 3266–3280.

[9] URL: https://github.com/google-research/pali, Accessed on: 15/04/2025, 10:00 GMT.