A Deep Learning Ensemble for Object Detection and Classification in Retail Stores

Miguel Fernandes
a21260869@isec.pt
Simão Paredes
sparedes@isec.pt
Ana Alves
aalves@isec.pt
Francisco Pereira
xico@isec.pt

Polytechnic of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes – Quinta da Nora, 3030-199 Coimbra, Portugal

Abstract

This paper presents a deep learning pipeline for object detection and classification of retail items. The aim is to develop an automated system to accurately identify and categorize products on store shelves. The proposed framework employs YOLOv7 for object detection, followed by classification using either convolutional neural networks (CNNs) or CLIP, a vision transformer model. Comparative experiments are conducted on a dataset of grocery product images. Metrics indicate that the YOLOv7 network effectively localizes individual items, while for classification CNNs exhibit reasonable but limited accuracy, while CLIP demonstrates stronger zero-shot generalization. The work demonstrates the feasibility of a robust vision-based product recognition system, while highlighting opportunities for performance optimization.

1 Introduction

Computer vision has transformed various industries, including healthcare, self-driving cars, security systems, and manufacturing. One crucial task in computer vision is the accurate detection and classification of distinct items under different scenarios. In retail scenarios, automated detection and recognition of products have diverse applications in stores, warehouses, and e-commerce, however, reliably identifying products in varied shelf configurations and imaging conditions is challenging.

This article focuses on a deep learning ensemble for object detection and classification in retail stores. The proposed system employs a pipeline with separate detection and classification stages, leveraging YOLOv7 for object detection and exploring two alternatives for the classification module. The first approach evaluates several pretrained CNN architectures fine-tuned on grocery product images, while the second applies CLIP, a model combining vision and language understanding via contrastive pretraining.

The paper subsequently proceeds with a review of related work in Section 2, an exposition of the proposed framework and selected datasets in Section 3, presentation and analysis of experimental results in Section 4 and conclusion and potential directions for future research in Section 5.

2 Related Work

Object detection and image classification play a crucial role in the automated understanding of retail environments. In recent years, deep learning methods have become the dominant approach for these tasks.

For object detection, two-stage models like Faster R-CNN were once prevalent, but one-stage detectors including YOLO and SSD now tend to be favored, as they offer speed without sacrificing accuracy [11, 13]. YOLOv7 is the latest iteration of the YOLO family, incorporating architecture improvements and new training strategies to boost performance [10].

In image classification, many different CNN architectures have demonstrated exceptional aptitude on vision tasks [1, 4, 7, 8]. However, their dependence on large labeled datasets poses challenges for real-world application. An alternative approach is contrastive self-supervised learning, where models like CLIP are pre-trained on unlabeled image-text pairs [2, 6, 12]. Fine-tuning CLIP on new classes using just textual descriptions enables zero-shot inference without collecting training data.

Prior works have applied deep learning for retail product recognition, but often rely solely on text or barcode reading [4, 7]. CNN's can provide more generalized visual understanding, as explored in systems developed in [8, 9].

The primary objective of the present work is to assess the framework's effectiveness in detecting and classifying items on retail shelves. The comparison of results between convolutional and transformer models aims to provide insights into their respective strengths and weaknesses. This analysis serves to gauge the framework's accuracy, efficiency, and applicability in real-world retail scenarios.

3 Methodology and Data

The proposed framework follows a two-step process (Figure 1).

1. Object Detection - YOLOv7 model to locate individual items.
2. Classification - Apply CNN's or CLIP on detected items.

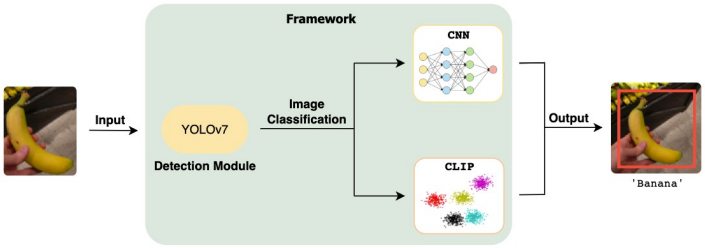

Figure 1: System pipeline diagram

3.1 Object Detection

For the object detection module , YOLOv7 was trained for 135 epochs (approximately 3 hours) and the training was performed with a batch size of 15, an image size of 640x640 and the Adam optimizer was used with a learning rate of 0.01. The model was trained on Google Colab using a Tesla T4 GPU with 16GB memory.

3.2 Image Classification

Image classification includes two sub-modules: Convolutional Neural Networks (CNNs) and Contrastive Language-Image Pretraining (CLIP). Using both allows comparison of traditional CNNs and novel vision transformers.

For the CNN's module, the models were trained on Kaggle using an Nvidia Tesla P100 GPU with 16GB memory. Three different convolutional neural network models were trained - ResNet101, EfficientNetB0, and MobileNetV2, providing a representative sample of modern convolutional networks. All models were first pretrained on ImageNet and then fine-tuned on the grocery dataset (described in 3.3). For coarse 3-class labeling, the models were trained for 10 epochs using a batch size of 32. The Adam optimizer was used with suitable learning rates for each model. Cross-entropy loss was optimized during training. For fine-grained 43-class labeling, early stopping with patience of 10 epochs was used. The models were trained for up to 50 epochs with the same batch size. Additional regularization techniques were not implemented.

CLIP leverages vision and language understanding, trained on image-text pairs to relate images and text. The openai/clip-vit-large-patch14 model is used in this work. CLIP's zero-shot approach removes the need for new training data when adding classes, enabling scalable classification.

3.3 Datasets

For object detection, the SKU110K dataset [3] was used, comprising more than 11 thousand shelf images with 1.7 million box annotations captured in densely packed scenarios. The Grocery Store Dataset [5] was employed for classification, containing around 5 thousand images with 81 fine-grained classes that are grouped into 42 coarse labels, which can further be categorized into 3 high-level groups: fruits, vegetables, and packaged goods, shot in an active store under diverse lighting and angles.

4 Results

YOLOv7 achieved a precision of 87.4%, recall of 81.2%,, and mAP@0.5 of 85.8%, on the test set. This confirms its capability to accurately detect shelf products despite diversity in scale, angle and appearance.

For image classification, Table 1 summarizes key classification metrics on the test set for coarse (3 classes) task.

Table 1: Training results with coarse labels.

Model	Train Loss	Train Accuracy	Test Loss	Test Accuracy	Training Time(s)
ResNet101	6.22%	97.70%	27.50%	91.40%	346.86
EfficientNetB0	5.21%	98.72%	9.01%	97.03%	274.36
MobileNetV2	12.52%	96.58%	29.43%	93.05%	275.48

For image classification, Table 2 summarizes key classification metrics on the test set for the fine-grained (43 classes) task.

Table 2: Training results with fine labels.

Model	Train Loss	Train Accuracy	Test Loss	Test Accuracy	Training Time(s)
ResNet101	28.95%	91.42%	6.05%	45.76%	1743.69
EfficientNetB0	0.56%	99.82%	8.34	56.60%	1352.44
MobileNetV2	4.27%	99.13%	8.03%	53.62%	1343.54

For coarse classification, all models exhibit >90% accuracy, with EfficientNetB0 achieving 97% on the test set. However, for fine-grained categorization, accuracies range between 45-57%, indicating significant room for improvement.

The complete pipeline was evaluated by first running object detection with YOLOv7, followed by image classification. For coarse classification, both EfficientNetB0 and CLIP showed strong performance, correctly classifying almost all detected objects. For fine classification, EfficientNetB0 struggled with misclassifications, while CLIP was more robust, properly classifying most items despite imperfect object detections. Overall, the results demonstrate promising capabilities of the object detection and image classification models for grocery product recognition. A result of the proposed pipeline can be seen in figure 4.

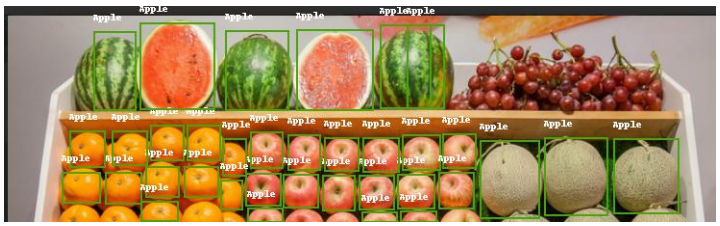

Figure 2: Classification module using CNN's

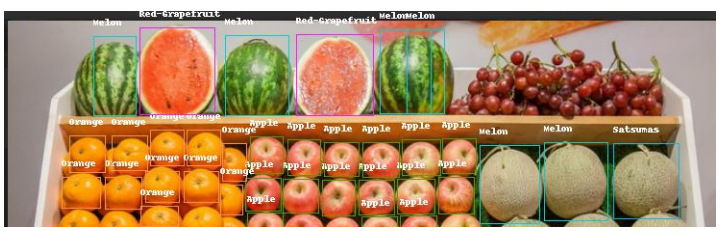

Figure 3: Classification module using CLIP

Figure 4: Results from both pipeline approaches after object detection for the fine label classification.

5 Conclusions

This work presented a deep learning pipeline for retail product detection and classification. Experiments demonstrate feasible classification between coarse product categories, but limitations in fine-grained recognition. The transformer-based CLIP model proved more adaptable than CNNs.

Priorities for future research include expanding the training data diversity, benchmarking on novel grocery item images, and investigating advanced self-supervised vision models. With additional tuning, the proposed approach could enable seamless integration of robust vision-driven product recognition in retail settings.

References

[1] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data 2021 8:1*, 8:1–74, 3 2021. ISSN 2196-1115. doi: 10.1186/S40537-021-00444-8.

[2] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, October 2021.

[3] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5222–5231, 6 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00537.

[4] Vânia Guimarães, Jéssica Nascimento, Paula Viana, and Pedro Carvalho. A review of recent advances and challenges in grocery label detection and recognition. *Applied Sciences*, 13(5), 2023. ISSN 2076-3417. doi: 10.3390/app13052871.

[5] Marcus Klasson, Cheng Zhang, and Hedvig Kjellström. A hierarchical grocery store image dataset with visual and semantic labels. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pages 491–500, 3 2019. doi: 10.1109/WACV.2019.00058.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th ICML*, volume 139, pages 8748–8763. PMLR, 18–24 Jul 2021.

[7] Prabu Selvam and Joseph Abraham Sundar Koilraj. A deep learning framework for grocery product detection and recognition. *Food Analytical Methods*, 15:3498–3522, 12 2022. ISSN 1936976X. doi: 10.1007/S12161-022-02384-2/FIGURES/15.

[8] Alessio Tonioni, Eugenio Serra, and Luigi Di Stefano. A deep learning pipeline for product recognition on store shelves. *IEEE 3rd International Conference on Image Processing, Applications and Systems, IPAS 2018*, pages 25–31, 7 2018. doi: 10.1109/IPAS.2018.8708890.

[9] Gül Varol and Rıdvan Salih Kuzu. Toward retail product recognition on grocery shelves. In Yulin Wang, Xudong Jiang, and David Zhang, editors, *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, volume 9443, page 944309. International Society for Optics and Photonics, SPIE, 2015. doi: 10.1117/12.2179127.

[10] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, June 2023.

[11] Xiongwei Wu, Doyen Sahoo, and Steven C.H. Hoi. Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64, 7 2020. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2020.01.085.

[12] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang

Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[13] Zhong Qiu Zhao, Peng Zheng, Shou Tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30:3212–3232, 11 2019. ISSN 21622388. doi: 10.1109/TNNLS.2018.2876865.