Université Polytechnique Hauts-de-France
INSA Hauts-de-France-Master 2 CaSSAD

Apprentissage Statistique Automatique II

TP Part 2

Project - Food item classification for automated retail sorting

Année 2025-2026

Raphaël SOUPART
Raphael.Soupart@uphf.fr

Younes CHAOUKI
younes.chaouki1@uphf.fr

Table des matières

1. Context and Objectives

As part of a major food retailer's supply chain automation initiative designed to mitigate operational costs, distribution delays, and waste caused by manual sorting errors, this report outlines the design of an image classification prototype for conveyor systems. Focusing on the strategic improvements implemented during the second session, the study contrasts a classic supervised approach (CNN), which is limited by the need for constant retraining, with a modern multimodal approach (VLM) capable of semantic understanding and "zero-shot" detection of new products.

The analysis of operational impacts demonstrates that, despite higher computational requirements, the adoption of VLMs offers superior agility regarding produce seasonality and drastically reduces system update lead times, thereby validating a more resilient and scalable deployment strategy.

2. Methodology

To meet the requirements, we explored two distinct architectures :

2.1. Approach A : Supervised Transfer Learning (MobileNetV2)

This approach targets performance and execution speed.

— **Model :** MobileNetV2 (pre-trained on ImageNet), chosen for its lightweight nature suitable for industrial constraints.

— **Improvement Strategy :**

  — **Data Augmentation :** Application of random rotations, zooms, and flips to simulate the random positioning of fruits on the belt.

  — **Fine-Tuning :** Unfreezing the last layers to adapt the filters to the specific textures of the fruits.

2.2. Approach B : Zero-Shot Learning (VLM SigLIP)

This exploratory approach uses a multimodal Large Language Model (Google SigLIP) to classify images without any specific training.

— **Method :** Semantic comparison between the image and textual descriptions.

— **Targeted Advantage :** Total flexibility to add new products to the catalog without retraining.

3. Results and Comparison

The tests were conducted on the *Fruits-360* dataset. The table below compares the performance obtained :

Metric	Baseline	MobileNetV2	VLM SigLIP
Architecture	Simple CNN	Transfer Learning	Vision-Language
Training	From Scratch	Fine-Tuning	None (Zero-Shot)
Accuracy	99% (overfitting)	96%	78%
Inference Speed	Very Fast	Very Fast	Slow

TABLE 1 – Performance comparison of the three models.

The supervised approach (MobileNetV2) significantly outperforms the VLM on the global task, gaining nearly 20 accuracy points.

4. Error Analysis and Operational Impact

The analysis of the VLM model's confusion matrix revealed critical weaknesses for an industrial application.

4.1. Technical Analysis

— **Semantic Confusion :** The VLM fails to distinguish visually similar fruits.
— **False Negatives :** A large portion of fruit is misclassified.

4.2. Business Impact

Deploying an imprecise model (VLM) would have direct consequences on the supply chain :

1. **Stock Errors :** Confusing varieties distorts inventory and leads to financial losses.
2. **Bottlenecks :** The high inference latency of the VLM is too slow for the rapid pace of a conveyor belt, risking a slowdown in distribution.
3. **Manual Rework :** The high error rate on certain classes would require costly human intervention to correct mistakes.

5. Conclusion and Recommendation

After testing both options, we strongly recommend using the MobileNetV2 (Fine-Tuned) model for the final prototype. This model is the best choice because it balances high accuracy **(over 95%)** with high speed. Speed is very important for our factory lines because we cannot afford to slow down the conveyor belts or cause delays in shipping.
On the other hand, the VLM approach is not ready for the main sorting line yet. Our tests showed that it is too slow for real-time work and often confuses fruits that look similar. Because of this, we should only use the VLM as a "backup" system to help identify new or strange items that the main model does not recognize.