CS3 Case Study: Fruit Image Classification with Deep Learning

Tamer El Aridi
DS 4002 – Spring 2025

Introduction: The Challenge of Automated Visual Recognition

In an increasingly data-driven world, the ability to automatically interpret and classify visual information is a cornerstone of modern Artificial Intelligence. From self-driving cars identifying pedestrians to medical systems detecting anomalies in scans, image classification has profound real-world applications. One fundamental yet illustrative domain for exploring these techniques is the classification of everyday objects, such as fruits.

Imagine you are a data scientist tasked with developing a system that can accurately identify different types of fruit from images. This capability could be used in automated supermarket checkouts, quality control in agricultural supply chains, or even dietary tracking applications.

The Scenario:

Your team has been provided with a substantial dataset of fruit images, encompassing various types like apples, bananas, and oranges, including different varieties within a single fruit category (e.g., multiple types of apples). The challenge is to leverage the power of deep learning, specifically Convolutional Neural Networks (CNNs), to build a robust image classification model. You will explore how these networks learn features from images and how techniques like transfer learning can be used to achieve high performance even with moderately sized datasets.

Objective:

The primary goal of this case study is to develop, train, and evaluate a Convolutional Neural Network capable of:

1. Accurately distinguishing between several distinct fruit categories.
2. Potentially identifying subtle differences between varieties of the same fruit (e.g., different apple types, if data allows).

3. Understanding the impact of data preprocessing and model architecture choices on classification performance.

Deliverable:

You will produce a concise technical report and a set of completed Jupyter Notebooks. Your report will detail your data exploration process, the model architecture, the training procedure, and a thorough evaluation of the model's performance on unseen data. This will involve analyzing key metrics such as accuracy, precision, recall, and confusion matrices to understand the model's strengths and weaknesses.

GitHub Repository:
A GitHub repository containing the necessary data (or links to it), starter Jupyter Notebooks for data exploration and model training, and relevant reference materials will be provided to guide your work.