Capstone Project 2

Sava Dashev

Fruit Classifier

Progress Report 1

Problem statement

Retail stores move ahead. Customers started scanning items for themselves. This pattern helps stores save money.

In grocery stores, customers buy fresh fruits. They do not always know the name of the fruit. To save time for customers and money we can use machine learning to automatically recognize types of fruit on the scale.

The goal is to program a classifier that will recognize type of fruit on the scale.

Dataset

The dataset was created for old competition. It is hosted on kaggle website.

The link:

[https://www.kaggle.com/chrisfilo/fruit-recognition](https://www.kaggle.com/chrisfilo/fruit-recognition#Carambola%20001003.png)

The size of the zipped data is about 8GB. When unzipped, it is a lot more.

I used Google Colab to host my notebooks. I initially downloaded data from kaggle, and saved it on my Google Drive. When I need the data, I upload it from the drive.

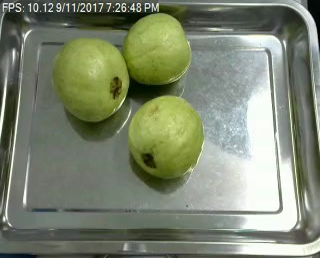
The unzipped data is in 15 folders, one for each class. Three of the classes have subfolders that contain images in subcategories. These are Apple, Guava and Kiwi. The other fruits do not have subfolders. All images are hosted in one folder.

Data cleaning

We checked the names of the images. Some of the images names contain “resized” and other contain “resized resized”. The resized images are of size (200, 200). These images are obtained from original images.

Most of the images are png files and rest are jpg files. One example is shown below.

Image '1GuavaуАВ2474.png':



The three resized images derived from this one: '1GuavaуАВ2474 resized.jpg',

'1GuavaуАВ2474 resized.png', '1GuavaуАВ2474 resized resized.png':

We see that these are three versions of the same image (note the time stamp is the same).

We decided to filter out these images.

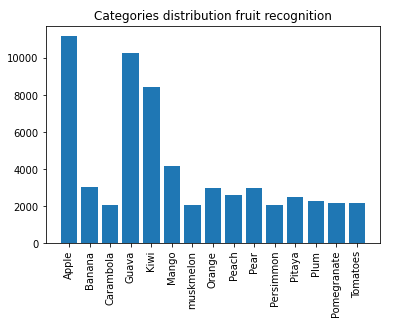
The rest of the images do not contain the string “resized”. We check the sizes of these images. It turns out that 6501 are size (480, 322) and 54633 are size (320, 258). The image 'Apple/Total Number of Apples/Apple 03100.jpg'

with size (320, 240).

The names and paths of the images are saved in file\_names.txt. To filter and investigate the images, we will use these paths.

We decided to use uniform size images to feed to classifier. We used the size (320, 258).

We will have only 15 classes. We count the images in the filtered list to see how many images are in each class.



We see that Apple has more than 10000 images, Guava and Kiwi have about 10 and 8 k images. All other classes have between about 4000 and 2000 images in the set we consider for machine learning.

The images were screened using the size of the image. This means none of the images is an empty file.

Preparation

Usually the images are placed in one directory when processed. We will use one size images and will not distort them. We will use the original zip file to save space and time.

We shuffle the paths of the files and corresponding classes. Next, we use train test split to create train and test sets for machine learning.

These variables are saved in files.

To feed the classifier, we will use the original zip file and the files with paths and y values for machine learning.

Next is to create classifiers.