COMP309 - Final Project
300467432 - W. Burroughs

Title of the report.
Your name.
Student number.
Course code (COMP309/AIML421).
Lecturer's name.
Assignment number/description.
Date (including the year).
Abstract

Abstract

Provide a concise summary of the entire report.
Include the project's objectives, methodology, major findings, and conclusions.

Introduction

Introduce the project and its context.
Clearly state the objectives of the project.
Explain the motivation behind the project.
Define the scope of the work.

Problem Investigation (Background)

Provide background information related to the problem you're addressing in your project.
Include relevant prior research and reading.
Highlight the importance of the problem.
Justify the need for your investigation.

Methodology

Exploratory Data Analysis (EDA)

Dataset
The dataset employed for this research project encompasses a total of 6,000 images, with a primary focus on classifying three distinct fruit types: tomato, cherry, and strawberry. These images exhibit a wide array of challenging characteristics, including variations in scale, colour, angles, background clutter, illumination, resolution, and the number of target items within each image. This dataset is divided into training and test sets, with 4,500 images allocated for training the deep Convolutional Neural Network (CNN) model, while the remaining 1,500 images are reserved for evaluation.

Data preprocessing was carried out, encompassing data cleansing and the resizing of most images to a uniform size, typically 300 x 300 pixels, to ensure consistent dimensions. It's imperative to note that the images are sourced from Flickr, and adherence to the respective terms of use is emphasised, restricting distribution beyond the course or public internet access. In summary, the dataset comprises 4,500 images of tomatoes, cherries, and strawberries, and it has been meticulously prepared to address the challenges posed by diverse image characteristics for subsequent deep CNN model development and evaluation.

Loading Data
In the preprocessing phase, the Orange Data Mining software, coupled with the Image Processing add-on, was utilised. To begin, the dataset, comprising 4,500 JPEG images, was imported into Orange using the "Import Images" module. Subsequently, the "Image Embedding" module was applied to process all images. This module harnesses Google's Inception v3 model, pre-trained on ImageNet, to generate image embeddings. These embeddings serve to transform the image data into a more convenient format, thereby facilitating streamlined downstream processing. The incorporation of the Inception v3 model and image embeddings enables the conversion of images into numerical representations, thus augmenting the effectiveness of subsequent machine learning tasks.

Initial Data Analysis of Features
The initial analysis of the dataset's features offers valuable insights into its characteristics. The dataset comprises a substantial number of features, with 2052 in total. These features exhibit variations in their distributions, as indicated by differences in mean values, modes, medians, and dispersions. The dispersion, represented by the range between the minimum and maximum values, varies across these numerous features, reflecting diverse data patterns.

One noteworthy aspect is the absence of missing values within the entire dataset. This comprehensive assessment of feature statistics forms the basis for subsequent data exploration, feature engineering, and model development within this data analysis project.

Data Preprocessing
The first step was to preprocess the dataset effectively, Orange was used to achieve this. Orange allowed for speedy preprocessing & aided in visualisation of the data.

The features *image* and *image name* were removed from consideration, as they bore no relevance to the target variable *category* and were essentially arbitrary names. Similarly, the features *height* and *width* were excluded, as they exhibited constant values across all instances and did not contribute meaningfully to the analysis. Additionally, the *size* feature was deemed non-essential for our study, as its relevance in the context of image classification was limited.

Since there was no missing data there was no need to impute any missing values.

Dimensionality Reduction
Feature Selection and Dimensionality Reduction played pivotal roles in shaping the model for this study. The Recursive Feature Elimination (RFE) technique was employed to systematically rank features based on their importance in the context of image classification. This process was guided by the RRelief (ReliefF) score, which assesses feature relevance and discrimination capabilities.

After a comprehensive evaluation of RRelief scores for all features, the decision was made to exclude those with RRelief values below 0.080. This criterion led to the selection of the top 716 features. By applying this approach, a balance was struck between data complexity and model accuracy, allowing for the streamlining of the dataset. Features contributing minimally to image classification were removed, while the most informative ones were retained, ultimately enhancing model performance.

Data Exportation
The resulting preprocessed data was exported from Orange and saved as a CSV file for subsequent processing using Python 3.

Model Building

Discuss any preprocessing steps you applied based on your EDA.

Baseline Model (Step 3)

Describe the baseline model you built (MLP) and how it was trained.
Explain the results and insights you gained from this baseline model.

Your Own CNN Model (Step 4)

Describe your customized CNN model.
Include details on the architecture, layers, and hyperparameters.

Tuning the CNN Model (Step 5)

Discuss any experiments and tuning you performed to improve the CNN model.
Explain the results of your tuning efforts.

Results and Discussion

Summarize the key takeaways from your project.
Compare the performance of the MLP and CNN models.

Discuss any significant differences and why they occurred.

Conclusions (and Future Work) ?

Provide your final conclusions about the project's outcomes.
Suggest possible future work or areas for improvement.

References

Include proper citations for any external sources, papers, or materials you referenced in your report.
Follow a consistent citation format (e.g., Harvard system).

Appendices

Include any supplementary material that supports your findings (e.g., additional charts, data, or code).
This structure aligns with the guidelines you provided while incorporating elements specific to your project's objectives and tasks. Make sure to keep your report clear, concise, and well-organized, and include appropriate visuals and evidence to support your findings and conclusions.