Plant Classifier and Information Retrieval System

SUBMITTED BY:

B HariCharan Goud - 12240360
G Srujan - 1224058
K Nithin -

Contents

1 Problem Statement

The "Plant Classifier and Information Retrieval System" aims to identify plant species on campus using images of leaves. By leveraging a manually curated dataset and advanced image classification techniques, the system predicts the most probable plant matches. Additionally, it retrieves key information from sources like Wikipedia to provide users with comprehensive knowledge about the identified plants. This solution intends to promote biodiversity awareness and simplify plant identification for students, researchers, and nature enthusiasts.

2 Motivation

Understanding and preserving campus biodiversity is crucial for fostering environmental awareness and ecological balance. However, identifying plant species manually can be time-consuming and requires botanical expertise. This project aims to bridge this gap by leveraging technology to make plant identification accessible and efficient. By combining image classification with information retrieval, the system empowers users to learn about diverse plant species effortlessly, promoting curiosity, education, and sustainability.

3 Methodology

3.1 Data collection and Augmentation:

For data collection, we captured approximately 20 plant species, with 7-10 images of leaves from each plant, including both front and back views. Each set of images was organized into folders labeled with the corresponding plant names. To enhance dataset diversity and size, data augmentation techniques were applied to each image, generating five new variations. These included transformations such as flipping, rotation, adjustments to brightness and contrast, and the addition of Gaussian noise and Gaussian blur. This augmented dataset ensures robustness and improves model performance during training.

3.2 Plant classification and detection:

The collected plant dataset of 19 classes is trained using ResNet50 model and the penultimate layer is modified to suit the 19 plant species classification task. In the training process Binary Cross-entropy loss was employed as the loss function, as this is typically used for classification tasks.
The model was evaluated on the testing dataset consisting of unseen plant images. The test accuracy of the model was 92.75%, indicating the model's high performance in classifying plant species correctly. Further on giving an input leaf image, the model computes a cosine similarity score between the input

image and every other plant image in the dataset.After calculating the similarity scores for all plants, the top 5 most similar plants are identified based on the highest similarity scores.

3.3 Information Retrieval:

This step focuses on fine-tuning Dense Passage Retrieval (DPR) models using plant-related data, including structured information from Wikipedia. The DPRQuestionEncoder and DPRContextEncoder are employed to encode user queries and passages, respectively, with Hugging Face tokenizers for preprocessing. Custom dense layers are integrated to refine the output embeddings for efficient similarity computation. The fine-tuning process optimizes the models to match user queries with relevant plant information, leveraging contrastive learning. The approach enables accurate semantic retrieval of plant descriptions, family classifications, and uses, facilitating precise information retrieval tailored to user inputs.

4 Results

In the image classification task, the training accuracy for the training and validation accuracy came as 95.95% and 93.95%. The test accuracy for the model was 94.49%. In the information retrieval part, the information of the top 5 most similar plants to a test leaf/plant are generated using finetuned DPR models.