Daphne Hong (30062705)
Matthew Ocando (30090876)

ENDG 511: Project Proposal

Defining the Problem

Cassava plants are the second-largest source of carbohydrates in Africa. They are key to food security as they can withstand harsh conditions. Unfortunately, viral diseases can significantly impact yield. This makes it important to identify diseases quickly so that the plants can be treated and spread can be limited. Currently, disease detection is quite labor-intensive and costly. Agricultural experts are hired to inspect the plants and diagnose them. Additionally, African farmers face constraints due to limited access to resources (e.g. mobile-quality cameras, low bandwidth etc.).

About the Dataset

A cassava leaf disease classification dataset was selected for this project (Cassava Leaf Disease Classification | Kaggle). This is a multi-class classification problem with the following classes: Cassava Bacterial Blight (CBB), Cassava Brown Streak Disease (CBSD), Cassava Green Mottle (CGM), Cassava Mosaic Disease (CMD), and Healthy. This dataset contains folders that hold the images for training and testing.

Additional files including the following:
- train.csv - contains image id and labels for classification
- label_num_to_disease_map.json - contains the class mapping for classification

Other files not listed are specific to the rules of the Kaggle competition that this dataset is associated with.

Methods

To identify cassava plants that have diseases, transfer learning will be used with pretrained weights from MobileNet v2. After establishing the baseline, the following model compression techniques will be used in differing magnitudes and combinations: image scaling, pruning, and quantization. Lastly, Grad-CAM (a visual interpretation technique) will be applied to analyze the features used by the model for classification.

Performance Metrics

Metrics such as model accuracy, model size, and a confusion matrix will be used to analyze the performance of this classifier. By comparing the performance metrics at varying levels, we can determine the optimal balance between accuracy, model size, and computational efficiency. These metrics are essential in understanding the trade-off between compression and performance, allowing us to make informed decisions about the deployment of the classifier.

Expected Analysis & Trade-Offs

Given the intended use-case for this model, we will be applying different compression techniques to analyze the trade-off between model size and accuracy. Finding the right balance between these two are crucial to ensure good model performance and a small size for deployment on constrained devices. Our expectation is to create graphs/diagrams to visualize the trade-offs that we anticipate, and by doing so hopefully reach informed conclusions regarding what might be considered an optimal model for deployment.

Daphne Hong (30062705)
Matthew Ocando (30090876)

Potential Challenges

It is our expectation that there will be two major challenges faced during the completion of this project. These are as follows:

1) *Training time*
   Given the limited resources available, it may become necessary to sacrifice certain aspects of training (e.g. number of training epochs, batch size, etc.). It is unfeasible to train these models for hours, limiting the extent of analysis for this project.

2) *Sparse Compression Intervals*
   Given the constraint on training resources, the intervals at which the compression techniques are explored may need to be rather sparse. For example, pruning may only be feasible to test with intervals of 20% given that a smaller interval would result in a vastly larger number of combinations across our three intended compression techniques and significantly increased training time.