Report on Cassava Disease Classification Project

Kaggle Team Name: GOST
Team Members

- Volviane Saphir Mfogo (smfogo@aimsammi.org)
- Gedeon Muhawenayo (gmuhawenayo@aimsammi.org)
- Sewade Olaolu Ogun (sogun@aimsammi.org)

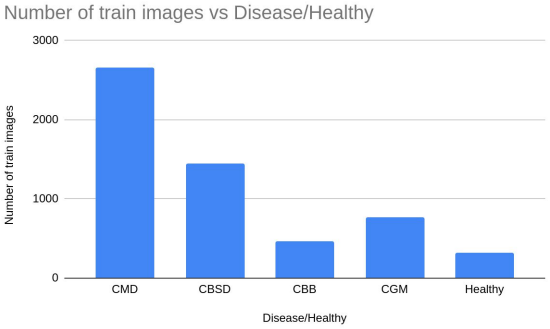
I. Introduction

The project involved classifying pictures of cassava leaves into 1 of 4 disease categories (or healthy). The disease categories include Cassava Mosaic Disease(CMD), Cassava Brown Streak Disease(CBSD), Cassava Bacterial Blight(CBB), Cassava Green Mite(CBB).

(a) Healthy (b) CBB (c) CGM (d) CMD (e) CBSD

Dataset

Dataset consists of **9,436** labeled and **12,595** unlabeled images of cassava plant leaves.



Statistics of the Cassava Dataset. Paper

The graph shows training examples for CMD is twice more than every of the other classes. This creates a class imbalance.

Project Assessment Criteria: Top-1 Accuracy on test set.

II. Approach

Discussion on Research Paper

We read the paper to understand how the data was generated and annotated. The dataset was crowdsourced from approximately 200 farmers in Uganda with the same mobile devices. We also observed that different diseases have their unique symptoms while observing the training set.

Data Preprocessing

- We split train data into 80% train and 20% validation using 5-fold stratified k-fold cross validation. The stratification was to cater for the imbalance in the dataset.
- Images were resized to 224px for small models and 500px for larger models.
- Different transforms such as centre cropping, rotation/affine transformation, horizontal and vertical flips.

III. Results and Experiment

Table 1: Table Showing Different Experiments Performed.

					Parameters						
No	Model	Train acc	Valid acc	Public LB	optimizer	Batch-size	Learning rate	Epochs	Drop-out	Folds	Extra-images
1	Resnet18	0.87	0.864	0.856	SGD	16	0.0002	22	No	5	No
3	Resnet34	0.8723	0.8736	0.8735	Adam	16	0.0002	13	No	2	No
4	Resnet50	0.8992	0.886	0.8741	SGD	16	0.001	9	No	5	No
5	Resnet50	0.9072	0.8993	0.90794	SGD	16	0.001	20	0.5	5	No
6	Resnext101	0.9266	0.906	0.912	SGD	16	0.0002	3	0.5	2	No
7	Resnext50	0.9651	0.909	**0.91655**	Adam	8	0.001	10	No	5	No
8	Ensemble of 5, 6, 7			**0.92185**							
9	Resnet50	0.9172	0.9108	0.910594	SGD	16	0.001	20	0.5	5	Yes
10	Ensemble of 5, 6, 9			**0.92384**							yes

What we learnt

- Cross validation is important to understand how the model generalizes, especially for small datasets like this one.
- Dropout can help to prevent overfitting, especially in models with large capacity such as ResNet50.

What did not work

- Extraimages did not significantly improve model performance during cross validation, when added to the training data.
- Some transforms like Random Erasing were not beneficial.

Dealing with Extra Images

- We used our top-3 performing models to perform a pseudo-labelling of the examples.
- If the 3 models agree on a class, we add the example to our training data.

Challenges

- Some images have more than one disease and could be classified into any of the classes. We tried label smoothing to cater for this.
- The large models took a lot of time to train, so difficult to iterate on ideas.