Maize Streak Virus

Kalema Martin[1], Nagaba Blessing[2], Namuli Sylvia[3], Ainomugisha Priscilla[4], and Ahwera Shifrah[5]

[1]Makerere Artificial Intelligence Laboratory

Maize streak virus (MSV) is a virus primarily known for causing maize streak disease (MSD) in its major host, and which also infects over 80 wild and domesticated grasses. MSV was first described by the South African entomologist Claude Fuller who referred to it in a 1901 report as "mealie variegation".

Keywords: Computer Vision, Maize Streak Disease, Transfer Learning, Artificial Intelligence

Introduction

Maize streak virus (MSV) occurs throughout Africa, where it causes what is probably the most serious viral crop disease on the continent. It is obligately transmitted by as many as six leafhopper species in the genus Cicadulina, but mainly by C. mbila and C. storeyi. In addition to maize, it can infect over 80 other species in the family Poaceae. Whereas 11 strains of MSV are currently known, only the MSV-A strain is known to cause economically significant streak disease in maize. Severe maize streak disease (MSD) manifests as pronounced, continuous parallel chlorotic streaks on leaves, with severe stunting of the affected plant and, usually, a failure to produce complete cobs or seed. Natural resistance to MSV in maize, and/or maize infections caused by non-maize-adapted MSV strains, can result in narrow, interrupted streaks and no obvious yield losses. MSV epidemiology is primarily governed by environmental influences on its vector species, resulting in erratic epidemics every 3–10 years. Even in epidemic years, disease incidences can vary from a few infected plants per field, with little associated yield loss, to 100% infection rates and complete yield loss.

Dataset

We utilized a publicly accessible dataset obtained from the Makerere Artificial Intelligence Lab for our maize streak virus (MSV) classification task. The dataset comprises 6489 images of maize plants exhibiting symptoms of MSV infection, along with corresponding labels indicating the presence or absence of the virus. These images were captured under diverse environmental conditions and at different growth stages of the maize plants.

Methodology

A. Data Preprocessing. The dataset was preprocessed to boost its diversity and robustness before training the classification model. These stages included scaling photographs to a standard resolution, normalizing the data to scale pixel values, and augmenting the dataset using rotation, flipping, and cropping.

B. MLOps pipeline. We integrated MLflow into our machine learning workflow for model registry/experiment tracking and management, while utilizing Streamlit for model deployment, creating a prediction service. This approach aligns with MLOps level 0, which represents a manual process for building and deploying ML models. MLflow facilitated the logging of hyperparameters, metrics, and artifacts associated with each experiment, providing a comprehensive overview of model training and evaluation. Meanwhile, Streamlit allowed us to develop a user-friendly interface for users to upload images of maize plants and obtain predictions on the presence of MSV. The manual process involved interactive steps, including data preparation, model training, and validation, often driven by experimental code executed in notebooks. Deployment of the trained model was solely focused on serving it as a prediction service, without active performance monitoring or frequent release iterations. CI and CD were not considered due to infrequent model version deployments.

C. Model Architecture. We used a pretrained VGG16 model, which is well-known for its ease of use and efficiency, especially in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Using several 3x3 convolutional filters stacked throughout the network, VGG16, with its 16 layers (13 convolutional layers and 3 fully connected layers), facilitates the learning of complicated hierarchical features. For downsampling, max-pooling layers with a 2x2 window and a stride of 2 are used, which decreases spatial dimensions and expands the receptive field. High-level feature aggregation is performed by fully connected layers, which operate as the classifier and introduce non-linearity through Rectified Linear Unit (ReLU) activation functions.

D. Model Training. The model was trained using a combination of transfer learning and fine-tuning techniques with a VGG16 architecture. Initially, a pre-trained VGG16 model, with weights learned from the ImageNet dataset, was utilized. We replaced the dense layers of the VGG16 model with new layers adapted for the specific task of maize streak virus (MSV) classification, effectively tuning the model to classify only two classes relevant to MSV. Specifically, the last fully connected layer was replaced with a new layer having two output units corresponding to the classes 'healthy' and 'maize streak disease'. Fine-tuning was then performed on our MSV dataset to adapt the model to this task. Stochastic gradient descent (SGD) optimizer with a learning rate scheduler and categorical cross-entropy loss function were used during the training process.

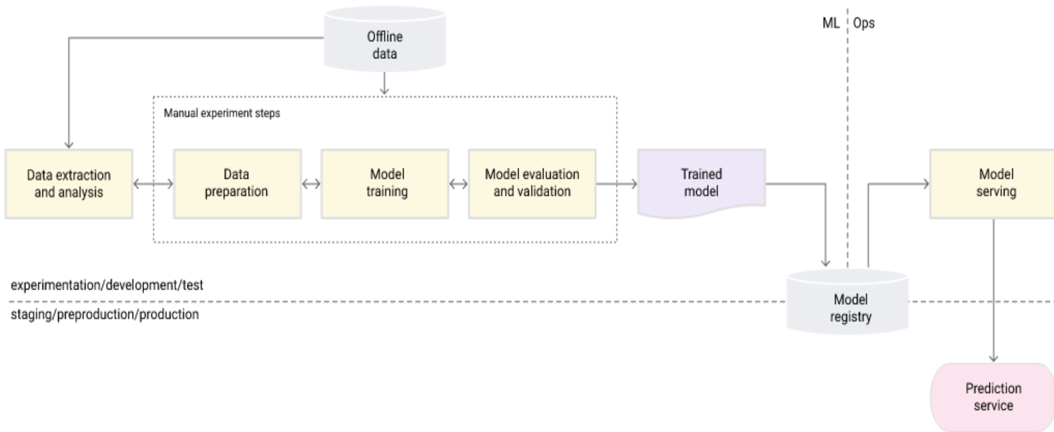

Fig. 1. MLOps Level 0

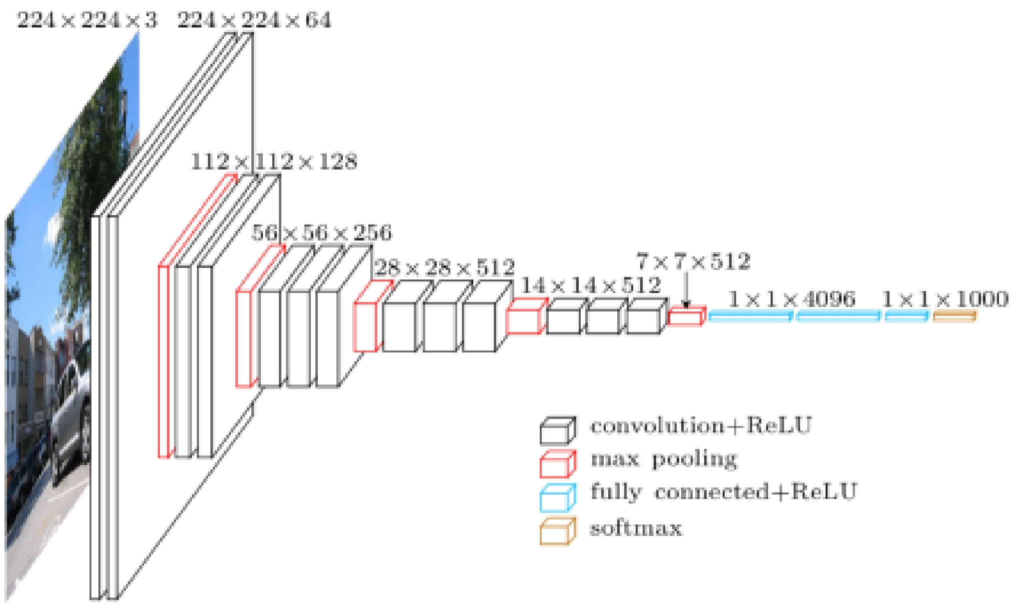

Fig. 2. VGGNet Architeture

Results

E. Performance. The performance of the trained VGG16 model was evaluated using accuracy. The best model achieved an accuracy of 95.42% and loss of 1.26 which indicates that the model is effective in classifying maize plants into 'healthy' and 'maize streak disease' categories based on the presence of maize streak virus (MSV).

F. Limitations. CNNs, like VGG16 in our study, are computationally expensive due to their complex architectures, demanding substantial computational resources for processing large image datasets. They often require large amounts of labeled training data to learn meaningful representations of features in the data. Insufficient training data may lead to overfitting, where the model performs well on the training data but generalizes poorly to unseen data. They are often considered black-box models, meaning it can be challenging to interpret and understand how the model makes predictions. This lack of interpretability may limit their use in applications where interpretability is crucial.

Future Work

We intend on deploying the prediction service into production that can handle expected workload and scale to accommodate increases in usage or data volume over time. The model/ prediction service will be installed on Heroku PAAS. We shall continuously monitor and evaluate the performance of the deployed CNN model in production. Collect feedback from users and stakeholders, and use this feedback to iteratively improve the model over time through retraining, finetuning, or updating the model architecture.

Conclusions

In summary, our study demonstrates the effectiveness of using a pretrained VGG16 model for maize streak virus (MSV) classification. Despite the computational expense, transfer learning and fine-tuning techniques yielded promising results in accurately detecting MSV-infected maize plants.

Links

You can visit the AI LAB at Makerere AI Lab. This dataset was reduced by selectiong 600 images from each class. dataset.