Title: Plant Disease Classification using Hybrid CNN + CAE Architecture to Reduce Computational Cost.

Who:
Alex Manioudakis - amanioud
Roshan Parikh - roshanparikh
Shruti Panse - spanse
Rio Wombacher - rwombach

Introduction: (Roshan) We want to make a model that can classify plants by their disease (or lack of) with as little parameters as possible to maximize efficiency.
Paper: https://www.sciencedirect.com/science/article/pii/S2589721721000180
DataSet (PlantVillage): https://www.kaggle.com/datasets/mohitsingh1804/plantvillage

- If you are implementing an existing paper, describe the paper's objectives and why you chose this paper.
    - We will be implementing Bedi, et al., which uses a hybrid deep learning system consisting of convolutional neural networks (CNNs) and convolutional auto-encoders (CAEs) to detect plant disease based on images. The reason for this hybrid system was to decrease computational strain while maintaining relatively high accuracy demonstrated in existing systems based only on CNNs. Specifically, this paper's algorithm was used to detect Bacterial Spot disease on peach plants. We will be implementing their paper on the dataset described below, which is more diverse. We chose this paper because of the hybrid algorithm specifically as well as the opportunity to apply this algorithm to more plant types and diseases, with the potential to determine if a generalized CNN and CAE program is applicable and accurate to a variety of plant disease scenarios.

- If you are doing something new, detail how you arrived at this topic and what motivated you.
    - N/A

- What kind of problem is this? Classification? Regression? Structured prediction? Reinforcement Learning? Unsupervised Learning? Etc.
    - This is a classification problem with supervised learning. Unlike the paper above, our dataset has 22 classes, which will complicate learning. Based on resource limitations, we may try to predict all 22 classes, reclassify the data into two classes (healthy and unhealthy), or both.

Related Work: Are you aware of any, or is there any prior work that you drew on to do your project? (Roshan)

- Please read and briefly summarize (no more than one paragraph) at least one paper/article/blog relevant to your topic beyond the paper you are re-implementing/novel idea you are researching.
    - N, et al. created a deep learning algorithm to detect diseased plants based on pictures of leaves. However, this study only used CNNs and no other methods (like CAEs). Their training dataset was also relatively small, with about 100 images, 10 of which were healthy with the rest unhealthy.
- In this section, also include URLs to any public implementations you find of the paper you're trying to implement. Please keep this as a "living list"–if you stumble across a new implementation later down the line, add it to this list.

Data: What data are you using (if any)? (Shruti)

- If you're using a standard dataset (e.g. MNIST), you can just mention that briefly. Otherwise, say something more about where your data come from (especially if there's anything interesting about how you will gather it).

https://www.kaggle.com/datasets/nirmalsankalana/crop-pest-and-disease-detection

We will be using the "Crop Pest and Disease Detection" data set from Kaggle. It is a crop pests/disease datasets sourced from local farms in Ghana.

- How big is it? Will you need to do significant preprocessing?

The dataset is presented in two folds; the raw images which consist of 24,881 images (6,549-Cashew, 7,508-Cassava, 5,389-Maize, and 5,435-Tomato) categorised into 22 classes. Preprocessing will need to be done, but the images are relatively standardized.

Methodology: What is the architecture of your model? (Rio)
We will be training our model with the data described above. We will take in the images of the types of plants and whether they are healthy or what disease they have andpreprocess them and then put them through a CAE. The CAE is used to obtain the compressed domain representation before classification to help us learn what are the most important features of the image. From there the outputs of the bottleneck layer in the CAE will be passed to the CNN classifier model which will output our predictions of whether the plant has a disease or not and if so what disease it is. We will be implementing a similar architecture to the paper and I think the hardest part of implementing the paper will be to achieve similar performance while no longer just doing binary classification of diseased or not but also predicting the type of disease which will allow the user to know possible remedies.

- How are you training the model?
- If you are implementing an existing paper, detail what you think will be the hardest part about implementing the model here.
- If you are doing something new, justify your design. Also note some backup ideas you may have to experiment with if you run into issues.

Metrics: What constitutes "success?"

- What experiments do you plan to run?
    - We're planning to primarily test accuracy
    - Also perhaps weighted F1 scores (for multiclass) to evaluate more granular performance
- For most of our assignments, we have looked at the accuracy of the model. Does the notion of "accuracy" apply for your project, or is some other metric more appropriate?\
    - Accuracy is appropriate here. It is a classification task
    - However, we are also more concerned with recall than precision due to the nature of our project. Since we are trying to help people understand the disease of their plants, we would want to air on a cautious side of things and not have too many false negatives.
- If you are implementing an existing project, detail what the authors of that paper were hoping to find and how they quantified the results of their model.
    - The authors are primarily concerned with the accuracy metric. They examine this for both training and testing.
    - They used RMSE to evaluate the loss
    - Furthermore, they are very aware of the number of parameters and use this to provide a basic understanding of the efficiency of the model. Importantly, their model achieved extremely high accuracy (>98%) with 9,914 parameters. Other competitive models used 10s-100s of millions!!
- What are your base, target, and stretch goals?
    - Base: 75% accuracy
    - Target: 85%-90% accuracy
    - Stretch: 99% accuracy

Ethics: Choose 2 of the following bullet points to discuss; not all questions will be relevant to all projects so try to pick questions where there's interesting engagement with your project. (Remember that there's not necessarily an ethical/unethical binary; rather, we want to encourage you to think critically about your problem setup.) (Roshan + Shruti)

- What broader societal issues are relevant to your chosen problem space?
    - Food production is an important domain for which this project applies. More accurate assessments of plant health, especially for agricultural produce, would

help prevent instability in food production, especially in countries with less agricultural output and more variable harvests.

- Why is Deep Learning a good approach to this problem?
    - Deep learning is a great fit for image-based problems like ours because it can pick up patterns in images without us having to manually choose what features to look at. In our project, we're combining a convolutional autoencoder (CAE) with a CNN. The CAE helps shrink the image down while keeping the most important parts, which makes things faster and less resource-heavy. Then the CNN takes that simplified version and figures out if the plant is healthy or what kind of disease it has. This setup gives us a nice balance between accuracy and efficiency, which is especially helpful if we want the model to run on low-power devices like smartphones in the field.
- What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative? What kind of underlying historical or societal biases might it contain?
- Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?
- How are you planning to quantify or measure error or success? What implications does your quantification have?
- Add your own: if there is an issue about your algorithm you would like to discuss or explain further, feel free to do so.

Division of labor: Briefly outline who will be responsible for which part(s) of the project.

- Preprocessing: Roshan, Shruti
    - Note: do a 70-20-10 split
    - Consider a binary prediction with more species
- Creating and Testing Layers: Rio, Alex
- Writeup/Discussion of Results: Everyone