AgroML: Few-shot Plant Disease Classification

EE 491: B.Tech. Project Stage I

Siddharth Khandelwal
Department of Electrical Engineering
Indian Institute of Technology Bombay
Email: 190070062@iitb.ac.in

Guided By Prof. Rajbabu Velmurugan

Abstract

Crop diseases is a major threat to food security in today's world, hence detection and classification of these diseases will play a key role in increasing agricultural produce. Automation of plant disease detection has proven to be useful by reducing the number of man hours needed and expertise in the field. Recent developments in computer vision in deep learning now make it possible to use automated systems to detect diseases in crops with high accuracy. However, developing such accurate models requires large datasets of plant diseases which are difficult to obtain and label. Rare and less documented diseases also will be difficult to classify using such models. Hence models that can detect plant disease while simultaneously requiring less labelled data are essential. We look into few-shot learning and approaches for it. We specially explore foundation model CLIP and whether it can be used for plant disease few-shot classification.

Contents

1 Literature Survey

The initial survey included looking into a specific plant disease - Bacterial Wilt disease and papers related to its classification and detection. Bacterial wilt disease is a wide spread disease which infects over 200 species of crops. It causes the wilting of leaves, and slowly whole of the plant to shrivel and die. In [1] the authors classify and grade the level of infection of bacterial wilt disease using a CNN and network after pre-processing and segmenting the image. Some papers also look into multi-spectral and multifractal analysis for detection of Bacterial wilt disease [2] [3]. Apart from images, other type of data can also be useful in detecting and classifying diseases. The authors in [4] use a image-text multi modal model tfor disease identification. Hyperspectral images can provide useful insights into disease growth and can also be used for plant disease detection [5]. Other modalities might include soil moisture data, temperature, rainfall, wind speed in the region can be of high significance for plan disease monitoring [6]. We look into few-shot learning and delve deeper into foundation model CLIP for few-shot and zero-shot classification of plant diseases.

2 Dataset

In this section we introduce the datasets used in the experiments shown later. There are mainly 2 datasets which have been used: Plant Village and Plant Doc dataset.

2.1 Plant Village

The Plant Village Dataset [7] which contains 39 classes, out of which 38 classes are of different healthy and disease infected plant leaves. One of the classes contains background images later removed from the dataset. Every image in the dataset is of the size 256 x 256. There are around 54000 total images in the dataset. The dataset contains single leaf images taken in a lab environment.

2.2 Plant Doc

The Plant Doc dataset is a more realistic dataset where the images have not been taken in a controlled enviornment but instead contains images from the field, internet etc. The images are varying in sizes which makes it tough for models to adapt to them. There are a total of 28 plant disease classes in it with a total of 2600 images.

3 Few-shot learning

Few-shot learning is a sub-area in machine learning where we aim to classify new data when you have only a few training samples with supervised information. In Few-shot learning we define a N-way-K-shot classification in which we have

1. Training Set :
   This consists of N class labels each with K labelled images (where K can be around 5 to 10)

2. Query Set :
   These contain the query images which need to be classified among the N classes

If the data is insufficient to constrain the problem, then one possible solution is to gain experience from other similar problems. To this end, most approaches characterize few-shot learning as a meta-learning problem.

3.1 Meta Learning

In this approach, we treat our N-way-K-shot classification problem as the TEST data and a similar large base dataset as the meta learning training set (TRAIN). In each training episode of the model we sample N classes and K support images per each classes along with Q query images from the large TRAIN set. The model is trained on this large dataset in episodes to maximize the accuracy on the Q query images. This way our model learns to classify unseen classes given K samples for each. We finally classify our less labelled dataset using this trained model in similar episodes. Figure 1 shows the setting in a meta learning approach.

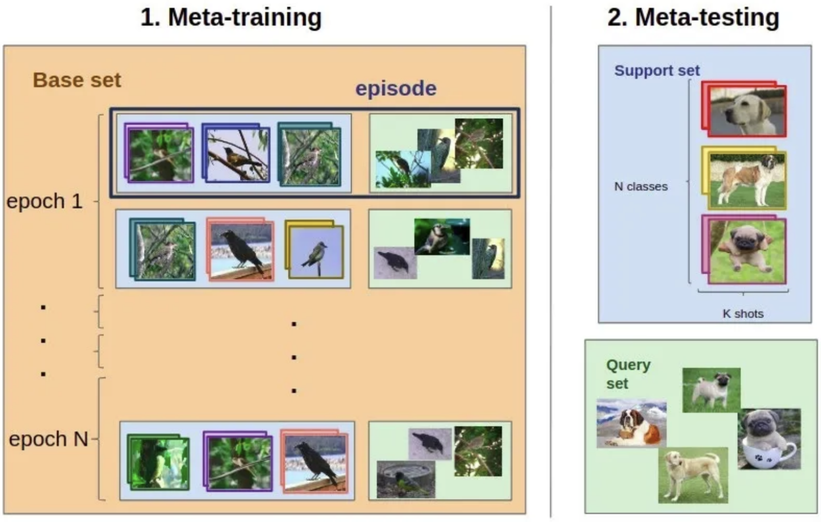

Figure 1: Meta learning [8]

4 CLIP

Foundation models in AI are models that are trained on a very broad set of datasets, and can be adopted for various application with minimal fine tuning. These contain billions of parameters which are pre-trained and available publicly. CLIP is an example of a foundation model.
CLIP model is trained on a wide variety of images along with a wide variety of natural language supervision that is abundantly available on the internet, hence the name Contrastive Language Image Pre-Training

(CLIP). As shown in Figure 2 , during training the model takes in a batch of N pairs of (image,text) input and learns a multi-modal embedding space by jointly training an image encoder and a text encoder to maximize the cosine similarity of the true N pairs and minimizing those of the rest possible pairs.

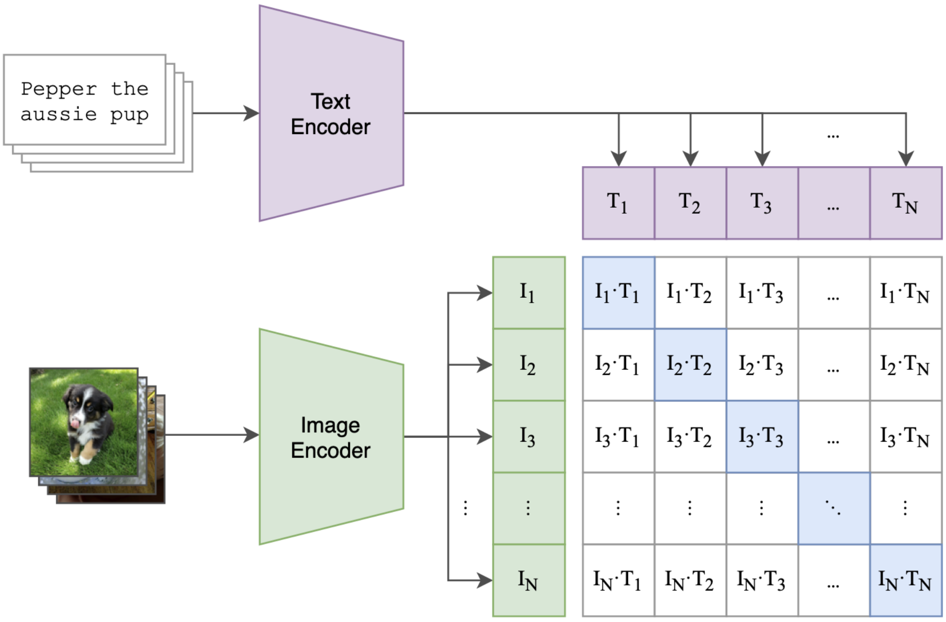
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Figure 2: CLIP model pre-training [9]

4.1 Zero shot with CLIP

In a zero shot setting we wish to classify an input image using a model which has not been trained on our dataset. The image is classified directly using the pre-trained CLIP model which is publicly available. The testing process of zero shot classification with CLIP is shown in Figure 3. For a particular image that needs to be classified, a number of classes are given in the prompt to the text encoder. The encoding which has the highest cosine similarity to the image encoding is selected as the predicted class.

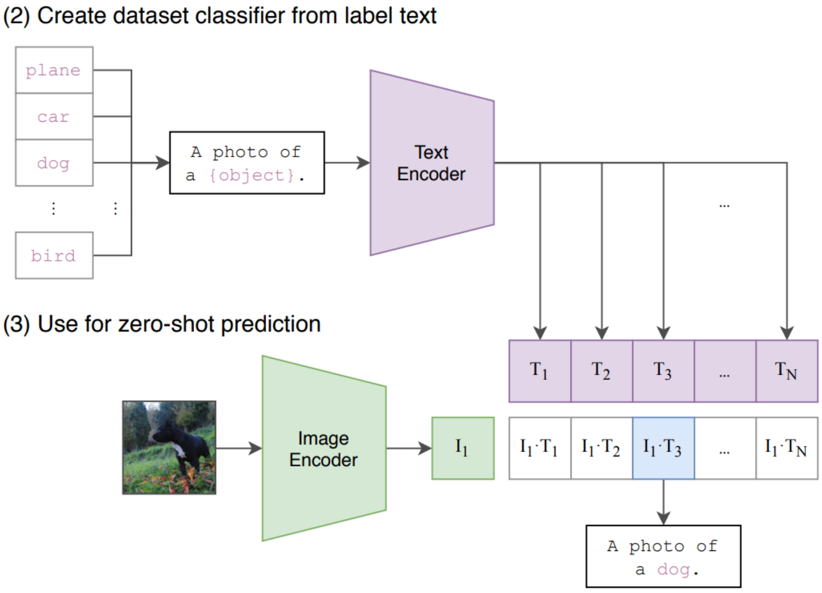

Figure 3: Zero shot with CLIP [9]

4.1.1 Experiment

We perform the experiment to differentiate between healthy and disease infected plant leaves using the Zero-shot learning. The input prompts include ”A photo of a healthy plant” and ”A photo of a infected plant”. The test set is then classified using these prompts. Achieved a binary classification accuracy of **85 %** on this task. This shows that CLIP model shows good zero shot performance to classify between disease infected and healthy leaves without any training on the Plant Village dataset.

Major problem in zero shot experiment with CLIP is that the prompts are difficult to engineer manually. The change in classification accuracy with even one word in the input text prompt are unpredictable and might be significant. Also we are unable to classify the particular disease in case of zero shot. To overcome these obstacles, the authors of [10] and [11] introduce methods to perform automatic learning of prompts.

5 CoOp

In Context Optimization(CoOp) manual prompt tuning by modeling context words with continuous vectors that are end-to-end learned from data while the massive pre-trained parameters are frozen is done. The context vectors or prompts are learnt through the back propagation of classification loss on the dataset. In Figure 4 we show how the possible classes are fed into the text encoder along with a prompt which is a continuous vector which can be trained to improve the classification accuracy.

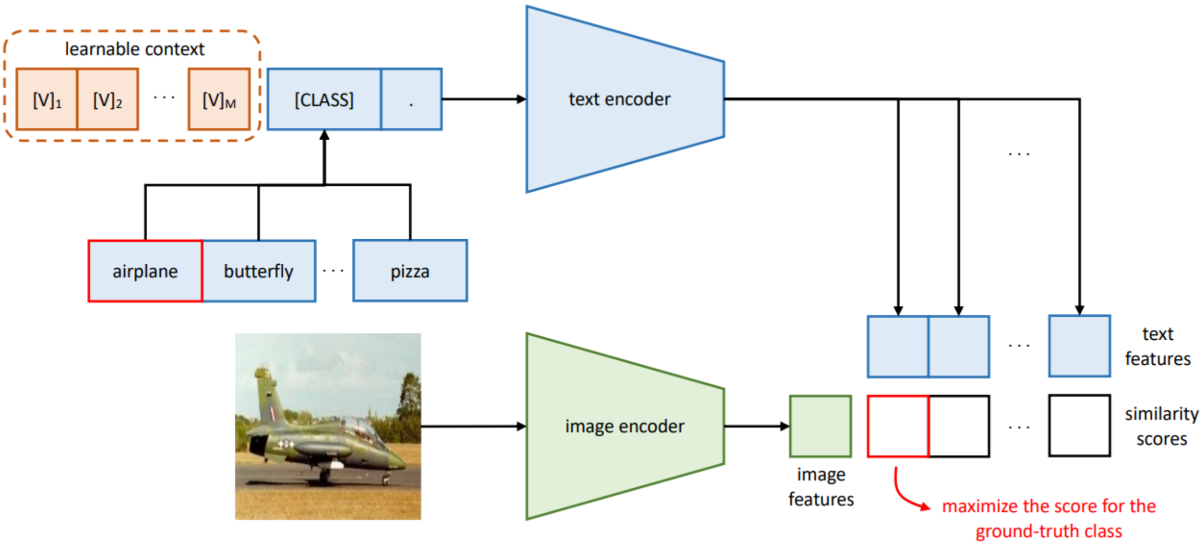

Figure 4: CoOp Model [10]

5.1 Few-shot Learning

The authors perform a few-shot learning experiment with this model on 11 famous datasets. In this setting, the model is trained on each classes containing N labelled samples each, The context vectors are learnt through the classification loss of these samples. Finally the model is tested on a test set keeping the context/prompt vectors fixed.

6 CoCoOp

The CoOp model is not generalizable to wider unseen classes within the same task. In CoCoOp which stands for Conditional Context Optimization, the prompts are optimized to characterize each instance which proves to be more robust against unseen classes. In this approach a light weight neural network called Meta-Net is learnt, to generate for each input image a conditional token (vector), which is then combined with the context vectors. This process is shown in Figure 5.

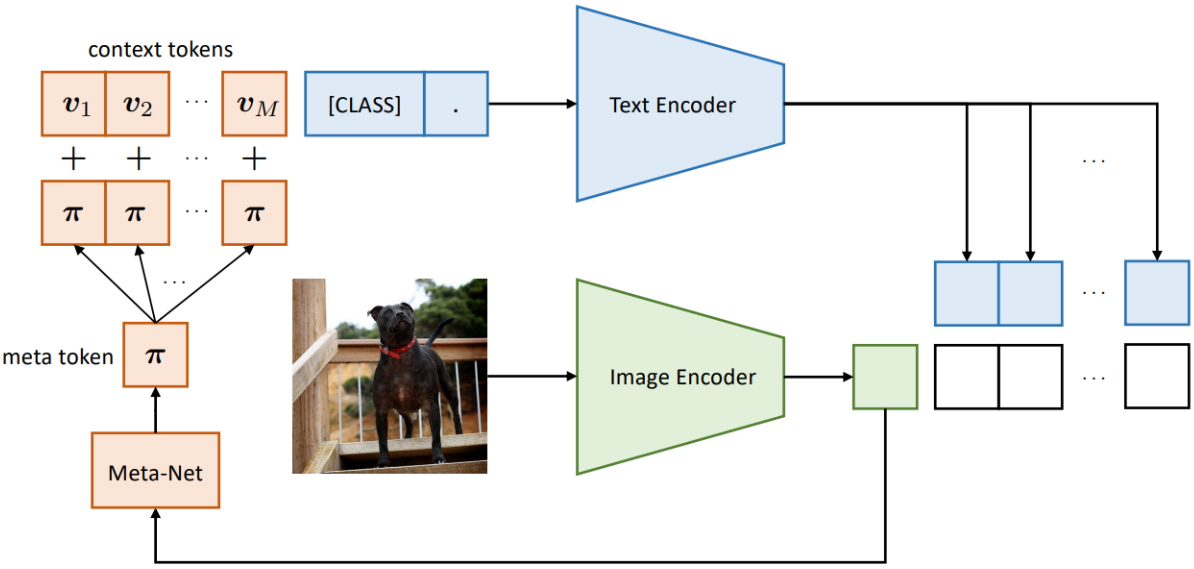

Figure 5: CoCoOp Model [11]

6.1 Generalization from Base to New Classes

In this experiment the dataset is split into 2 groups, one as base classes and the other as new classes. Learning-based models, i.e., CoOp and CoCoOp, are trained using only the base classes in an N-way-K-shot fashion, while evaluation is conducted on the base and new classes separately to test generalizability.

7 Experiments

All the experiments were performed in python using the pytorch framework. The training was done on a NVIDIA GeForce GTX 1080 Ti.

7.1 Prototypical Network Few-shot Classification

In this experiment we perform a N-way-K-shot classification on the Plant Doc dataset using the prototypical network model [12]. The Plant Doc dataset is divided into 19 base classes and 6 new or test classes. The new or test classes contain only 20 examples each. We train the prototypical network on the large base set and finally observe the classification accuracy on the unseen classes of the test set. The results are shown in the Table 1. We comapre these accuracy numbers with those obtained on the Plant Village dataset. We observe a large drop in accuracy numbers in the Plant Doc dataset since this datasets consists of complex images taken in an uncontrolled environment with varying noisy backgrounds etc. Hence the model is not able to learn to classify properly.

	Plant Village		Plant Doc	
N-way-K-shot	Training Accuracy (%)	Test Accuracy (%)	Training Accuracy (%)	Test Accuracy (%)
5-way-5-shot	93	82.8	82	46
5-way-10-shot	96	87	85	48

Table 1: Prototypical Few-shot Classification

7.2 Linear Probe with CLIP

Linear probing involves training a simple logistic regression on image features obtained from the CLIP's pre-trained image encoder. The dataset used for this experiment is the Plant Village dataset. The training set consists of all the 38 classes with N examples each in case of N-shot. The test set contains the same 38 classes with each class having 100-300 examples each. The Table 2 shows the classification accuracy results on the test set using linear probe.

	1 shot	2 shot	4 shot	8 shot	16 shot
Accuracy (%)	58.8	76.3	87	91.4	94.2

Table 2: Linear Probe Classification Accuracy

7.3 Few-shot learning using CoOp

In this setting instead of linear probe we test the CoOp model in the few-shot learning approach. The dataset used for this experiment is the Plant Village dataset. The training set consists of all the 38 classes with N examples each in case of N-shot. The test set contains the same 38 classes with each class having 100-300 examples each. The Table 3 shows the classification accuracy results on the test set using linear probe.

	1 shot	2 shot	4 shot	8 shot	16 shot
Accuracy (%)	27 ±3	50	53 ±5	74 ±2	82

Table 3: CoOp Few-shot Learning Accuracy

7.4 Generalization to Unseen Classes

The CoOp and CoCoOp models are trained on 30 classes out of the 38 classes in the Plant Village dataset. The rest 8 classes are used for evaluating the model and its performance on unseen classes. These 8 classes consists of 20 examples each. The CoCoOp model is relatively heavier to run, and hence the training accuracies might be increased using a bigger GPU with larger memory. The Table 4 shpws the various accuracy results.

	CoOp		CoCoOp	
K-shot	Training Accuracy (%)	Test Accuracy (%)	Training Accuracy (%)	Test Accuracy (%)
5-shot	78	26	58	38
10-shot	86	13	66	20

Table 4: Generalization to Unseen Classes on Plant Village

8 Conclusion

We do not observe desirable results on few shot classification with CoOp and CoCoOp models. The accuracy on the unseen classes are quite low and cannot be expected to be used in the real field. The reason behind this might be due to the CLIP model being trained on a everyday object images whereas, the plant village dataset consist of images of leaf with difference due to spots on the leaf, shape of the leaf etc. The model might not be able to adapt to such images using only N images (where N = 16).

Hence we see that learning with limited examples per class with Linear Probe and using CoOp is feasible for the Plant Village dataset. Whereas for generalization to unseen classes in the Plant Village dataset do not give good accuracy results.

9 Future Work

The CoOp and CoCoOp models might give better results if the text prompt describes the spots, patterns, or shape of leaf so that each prompt is better tuned for each class. Future work might involve looking into this.

References

[1] D. Ashebir and G. Tadesse, “Bwenet: Detection and grading of bacterial wilt using deep convolutional neural network,” *Indian Journal of Science and Technology*, vol. 15, no. 22, pp. 1100–1111, 2022.

[2] P. Chávez Dulanto, C. Yarleque, H. Loayza, V. Mares, P. Hancco, S. Priou, M. Márquez, P. Adolfo, P. Zorogastua, J. Flexas, and Q. Roberto, “Detection of bacterial wilt infection caused by ralstonia solanacearum in potato (solanum tuberosum l.) through multifractal analysis applied to remotely sensed data,” *Precision Agriculture*, vol. 13, pp. 236–255, 04 2011.

[3] Y. Cen, Y. Huang, S. Hu, L. Zhang, and J. Zhang, “Early detection of bacterial wilt in tomato with portable hyperspectral spectrometer,” *Remote Sensing*, vol. 14, no. 12, 2022.

[4] J. Zhou, J. Li, C. Wang, H. Wu, C. Zhao, and G. Teng, “Crop disease identification and interpretation method based on multimodal deep learning,” *Computers and Electronics in Agriculture*, vol. 189, p. 106408, 2021.

[5] P. Moghadam, D. Ward, E. Goan, S. Jayawardena, P. Sikka, and E. Hernandez, “Plant disease detection using hyperspectral imaging,” in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8, 2017.

[6] M. Kumar, A. Kumar, and V. S. Palaparthy, “Soil sensors-based prediction system for plant diseases using exploratory data analysis and machine learning,” *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17455–17468, 2021.

[7] G. J, ARUN PANDIAN; GOPAL, “Data for: Identification of plant leaf diseases using a 9-layer deep convolutional neural network,” 2019.

[8] E. Bennequin, “Few-shot image classification with meta-learning,”

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, pp. 2337–2348, jul 2022.

[11] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” 2022.

[12] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” *CoRR*, vol. abs/1703.05175, 2017.