*Semi-supervised Leaf Pathology Classification*

Singh, Manmeet  
Charles W. Davidson College of Engineering  
San Jose State UniversitySan Jose, California  
manmeet.singh01@sjsu.edu

Vadgama, Chinmay  
Charles W. Davidson College of Engineering  
San Jose State UniversitySan Jose, California  
[chinmay.vadgama@sjsu.edu](mailto:chinmay.vadgama@sjsu.edu)

Asha Balshiram Aher  
Charles W. Davidson College of Engineering  
San Jose State UniversitySan Jose, Californiaashabalshiram.aher@sjsu.edu  
  
  
Ronak Mehta  
Charles W. Davidson College of Engineering  
San Jose State UniversitySan Jose, Californiaronak.mehta@sjsu.edu

***Abstract*—This e**

***Keywords—component, formatting, style, styling, insert (****key words****)***

# Introduction

Misdiagnosis of the many diseases impacting agricultural crops can lead to misuse of chemicals leading to the emergence of resistant pathogen strains, increased input costs, and more outbreaks with significant economic loss and environmental impacts. Current disease diagnosis based on human scouting is time-consuming and expensive, and although computer-vision based models have the promise to increase efficiency, the great variance in symptoms due to age of infected tissues, genetic variations, and light conditions within trees decreases the accuracy of detection.

We explore plant disease detection by following computer vision based deep learning architectures on a large plant image dataset. We explore traditional supervised architectures such as ResNet-50, InceptionResNetV2 etc. and compare it with new self-supervised learning approaches such as SimCLR architectures. Generally, there is very little labeled dataset available for plants and specific diseases. Using self-supervised techniques such as SimCLR could provide ground for using large unlabeled dataset on top of learning from small labeled dataset. New data can be collected with help of drone-based cameras without any manual help and that data could be used as an unlabeled source for SimCLR network. This could get us more generalization and model adaptation can be increased even if we don’t have higher accuracy than supervised techniques.

# Related Work [Ronak]

# Data

Initially, we were exploring the medical domain and chose Alzheimer disease detection. The Alzheimer classification was carried out on the OASIS-1 Alzheimer as well as Alzheimer’s severity dataset. OASIS-1 is a cross-sectional MRI data in young, middle aged, non-demented and demented older adults. OASIS 1 has 416 subjects with 434 MRI scans. We carried out data engineering and data augmentation on this dataset and implemented SimCLR. There was a scope for accuracy improvement.

We carried out our classification on Plant disease detection. Two datasets were used in order to carry out the experiment. a) Plant village b) Plant pathology.

Plant pathology dataset has images of apple leaves and those are distinguished between healthy, scab & rust. Plant pathology has a total 3642 samples of above combination leaves. Plant pathology dataset contains different plant leaves including tomato, potato and bell paper.

# Methods

For the semi-supervised learning approach, we used simclrv2 which builds on top of contrastive learning based simclr. The contrastive approaches use heavy data augmentation followed by learning of the nonlinear transformation between representation and the contrastive loss which substantially improves the quality of learned representations. Lastly, the contrastive learning benefits from larger batch sizes compared to supervised learning. By combining these methodologies, we are able to achieve high performance with small amounts of labeled data.

Specific to simclrv2, a large amount of unlabeled data is used for unsupervised pre-training followed by supervised fine-tuning. This paradigm uses unlabeled data in a task-agnostic way initially. The key ingredient of this approach are the deep and wide networks during pretraining and fine-tuning. The fewer the labels in this approach, the more the pre-training benefits from this approach.

In our first approach, we used simclrv2 to perform unsupervised pre training on the OASIS-1 dataset. Next, we fine-tuned using the Alzheimer's severity dataset from Kaggle. After exhausting the possibilities for this approach, we pivoted to pre-training with plant village dataset and fine-tuned on the plant pathology dataset.

We trained our models on a combination of using the P100 GPU on the HPC machine, as well as the Google Colab environment.

# Experiments

Initially, we used SimCLRv2 to pretrain on the OASIS-1 dataset. The idea was to build an embedding space that could then be used to finetune on the Alzheimer’s severity dataset. This dataset included images from 4 classes ranging from normal to high severity.

After pre-training, we did not receive great accuracy during this methodology, likely related to the fact that we need a huge pre-training dataset to gain real benefits of the contrastive approach.

Next, we tried to pre-train on plant village dataset which is a multiclass classification dataset with 38 classes. Each class contains the leaf type as well as the disease/condition being experienced by the leaf. We achieved better accuracy after fine-tuning the plant village dataset on plant pathology dataset. We followed the standard procedure for finetuning where we removed the first projection head, attached a fine-tuning layer to the pretrained model and backpropagated the loss for this dataset. We were about to achieve 85% accuracy with this approach.

# Conclusion

In this paper, we study how to develop contrastive learning of visual representations methods to accurately detect plant disease from plant leaf images. For the ease of open research in this area, we build SimCLR model using plant pathology dataset which contains leaf images with different combination of the leaf situations which includes healthy, scab & rust. This dataset is publicly available to the date. In Deep learning models, overfitting is high risk for data-hungry deep learning models. In order to tackle this problem, we develop a model which is data efficient and is able to alleviate data deficiency. In this paper we propose an approach of a simple framework for contrastive learning of visual representations that learns expressive and unbiased visual feature representations that are robust to overfitting. In order to show effectiveness of our methods, we carry out extensive experiments.

##### References

1. Ranjita Thapa1 , Noah Snavely2 , Serge Belongie2 , Awais Khan\*1,”The Plant Pathology 2020 challenge dataset to classify foliar disease of apples” arXiv:2004.11958