Plant Seedling Classification

Yi Ding, Zhuoran Wu, Yu Xiao

GEORGETOWN UNIVERSITY 17 89

Abstract

For several decades, researchers have been working on systems that aim to perform site-specific weed control. Although several prototypes and case studies showed promising results, a truly accurate and reliable systems is still yet to come. Plant Seedlings Classification is a kaggle competition to find the best classification model that contributes to such systems. Plant Seedlings Classification is a typical supervised, multiclass classification task, with the goal to classify 12 different weeds in a crop seedling image dataset released by the Arahus University. This dataset contains approximately 960 unique plants belonging to 12 species at several growth stages. We applied both transfer learning and customized CNN models, and achieved near perfect micro-averaged F1 scores. In addition to models, we also experimented with customized loss function for potential improvement.

Introduction

Plant seedling dataset contains separated train and test dataset. The train set has 12 sub-folders with images of 12 species, which could be converted to labelled dataset. The test data set is a set of unlabelled images. Prediction results are evaluated by micro-averaged F1 scores. The goal of the project is to accurately classify 12 species. We mainly adopted CNN models in this project. We started by fine-tuning pretrained models such as VGG19 and Xception. Then, we experimented with several customized CNN models and we found customized models could achieve better classification results. Furthermore, we designed a customized loss function that converges faster than regular categorical cross-entropy loss function and reduced training time. There are a few limitations of this dataset that posed some challenges. For example, the plant seedling dataset is imbalanced. In addition, most images have backgrounds including barcode, pebblestone, metal device and wall. We used multiplied data augmentations and background blurry methods to alleviate the influences.

Methods

CNN: Transfer Learning: (pre-trained model)

We have tried several different pre-trained models and concat those model in several different ways.

- **VGG 19**
- Freezed first 5 layers and trained the rest of the layers with data augmentation. Also experimented with balanced dataset.
- **DenseNet 201, Xception, Inception V3**
  Used pre-trained models as feature selector. Also experimented with freezing more layers.

Methods

- **Customized Loss Function**
- **Data Cleaning & Balance Data**
  - **Adaptive synthetic sampling approach for imbalanced learning**
  - **Synthetic Minority Over-sampling Technique**
- **Data Augmentation**
  - **Scaling, Cropping, Flipping, Rotation, Translation, Adding Noise, Changing lighting conditions**
- **Snapshot Ensemble**
  - **Take snapshots of weights and biases and ensemble these networks**

Results & Discussion

Results

1. Pretrained model

We chose VGG19 as the baseline model. We Freezed the first 5 layers of the pre-trained VGG19 model and trained the rest of the layers. The model achieved F1 score 0.79722. With the balanced dataset (resampling) with 500 images for each class, F1 score slightly dropped to 0.76826. For pre-trained models with more layers (such as DenseNet 201), freezing first few layers resulted in extremely low performance.

2. Customized CNN

We apply different CNN model from scratch.

Currently, we apply a neural network with 9 layers and 9 batch normalization layers with 1772516 trainable parameters. The result of the public score is 0.97103. Then we applied all the methods such as data balance, data augmentation and then we could reach 0.9798 and 0.9823

3. Customized Loss Function

We are trying to combine the two most similar class as a superclass. And construct a tree-like hierarchical loss function base on the softmax.

In our experiment, the accuracy of our new loss function is not higher than the previous loss function. However, we find out that when the image number of one of class in the super-class is relatively low, our new loss function may have a chance to beat the original loss function.

Figures

Following are sample images for each species:

Fig. 1 Example images of 12 species.

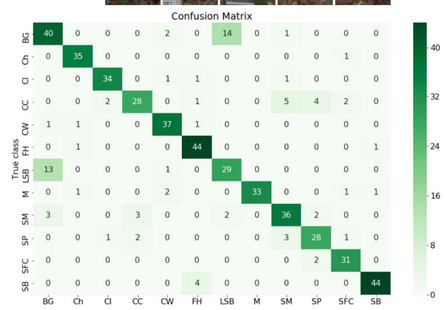

Fig. 2. True class vs Predicted class in a Confusion Matrix [1]

Conclusions

This work proposed plant seedling classification methods based on CNN. We also applied several different pre and post possessing methods such as data augmentation and ensemble method. Furthermore, we proved and applied customized Loss function for better learning the model. The final results shows that CNN is a great tool for image classification and we could reach 0.98-0.99 F1 score.

Reference

[1] Kaggle #1 Winning Approach for Image Classification Challenge

[2] Giselsson, Thomas Mosgaard, Dyrmann, Mads, J\orgensen, Rasmus Nyholm, Jensen, Peter Kryger & Midtiby, Henrik Skov (2017). A Public Image Database for Benchmark of Plant Seedling Classification Algorithms.

[3] Wu, Cinna, Mark Tygert, and Yann LeCun. "Hierarchical loss for classification." arXiv preprint arXiv:1709.01062 (2017).

Poster produced by Faculty & Curriculum Support, Georgetown University School of Medicine