CS Capstone Problem Statement

October 20, 2017

CS461 Fall 2017

Real-time Seed Identification

Prepared for

Oregon State University Crop Science Department

Daniel Curry

Prepared by

Group 11

Green Team

Omar Elgebaly
Danny Yang
Vinayaka Thompson

Abstract

The purpose of this project is to develop a solution that will accurately differentiate between desired grass seed and unwanted grass seed. The OSU Crop Science department are given batches of 25,000 seeds at a time. These batches are received from various clients such as golf courses, gardens, etc. When the Crop Science departments receive these seeds, there are off-types embedded within those packets that need to be removed. The reasoning behind this is because the off-types can contaminate the batch and cause weeds and other undesirable plants to sprout. The Crop Science departments job is to pull out all the off-type seeds by hand through visual inspection. The seeds are magnified and the analysts differentiate between the good seed and the bad seed by looking at different markers such as shape, texture, color, etc. This is an extremely time-consuming process that is tedious and causes eyestrain, headaches, and backaches. Our job is to utilize a machine learning tool that can identify the off-type seeds as off-types so they can be removed from the batch. The challenge is that the classifier must be able to remove every off-type. According to our client, the batches are generally 99 percent desired type and 1 percent off-type. This means we need to implement an algorithm with a relatively low margin of error that will accurately identify all the off-types.

Contents

1 Definition and Description of Problem

The OSU Crop Science Department takes batches of 25,000 grass seed at a time from clients. The department is paid to sort through the seed and separate the desired seed from the off-types, which will unavoidably be mixed in. The reasoning behind taking out these off-types is because a client such as a golf course owner will desire a specific type of grass seed to be planted. If the off-types are planted with the grass seed, it could potentially result in contamination and the spreading of weeds. The Crop Science Department has analysts that perform this sorting job by hand. They sit for hours a day and differentiate between the off-types and the desired seed by looking at texture, shape, and color with the aid of a high-powered lens. This is a tedious process that causes eyestrain, headaches, and backaches. It is also inefficient because it is not only slow but also prone to human error, especially after sifting through seeds for long periods of time.

2 Proposed Solution

Our proposed solution involves using a high-powered lens camera to capture and train images in a classifier. First, we will capture 25,000 images of the desired seed, Tall Fescue. The next step is to normalize the data in such a way that it reduces redundancy and getting all the data on the same scale. The goal here is to train the images with a margin of error of 25 percent. This means when the algorithm takes in an image of Tall Fescue seed as input, it should do nothing. When the algorithm takes in an image of an off-type grass seed as input, it should flag it as an off-type.

3 Performance Metrics

A complete version of this project would ideally be able to identify seeds of the desired grass type in real-time or more accurately identify what is not the desired grass seed type with a 25In addition to having a high accuracy, the software needs to be able to identify the seeds in real time. This means the testing phase of the machine learning algorithm should be able to accurately identify and flag off-types. Everything beyond that is up to the MIME group that is creating the conveyor belt and hooking it up to the computer. There are many potential challenges with this project. The first challenge I see is that we may not have the resources to feasibly get a perfect accuracy. Between our group, we have one Nvidia GPU. Even with that, training a batch of 25,000 images could take up to 6 days. We need to figure out exactly what kind of requirements can be expected with the hardware we have. Another notable challenge I see is the speed at which the classifier will be able to test the real-time images of the seeds being inputted. Our client expects to have a conveyor belt, with a camera suspended in the middle. It is hooked up to a computer and constantly uploading images of seeds to be tested. If the classifier classified the seed as an off-type, it is flagged and the computer will send a signal to the conveyor belt, which will indicate it should stop. Although the conveyor belt is not our responsibility, the testing phase needs to be quick enough that it can keep up with the belt. Finally, one of the biggest challenges is in the data collection phase. The grass seed is incredibly small, which means that a camera with a high-powered lens is required. Any visual noise could drastically affect accuracy. Furthermore, the normalization of the data could prove problematic if our images are not sufficiently clear. It is imperative we choose the right machine learning tool for this project as we do not have the time to spend a week training. Tensorflow is an attractive candidate for this job but it is implemented in Python, which leads me to believe that training could take too long to be feasible. Finding the right tool is imperative as we need to something that can process and train our images in a feasible window of time if we are to get accurate results.