Capstone Project 1: Final Report

Name: Jaime Ortiz

1. Introduction

Plants are the most intensively studied taxa in invasion sciences (Pyšek et al. 2008). Nevertheless, programs to control invasive plants have had very limited success (Buddenhagen et al. 2004). Mechanical and chemical approaches to control invasive plants are common techniques implemented to deal with these species (Kettenring and Adams 2011). Managing introduced species entails the allocation of usually scarce resources by setting priorities for control for particular species and deciding which methods will be cost efficient (Maguire 2004). For instance, tangles of the highly invasive plant kudzu overwhelm trees in Georgia while cane toads threaten habitats in over a dozen countries worldwide. These are just two invasive species of many which can have damaging effects on the environment, the economy, and even human health. Despite widespread impact, efforts to track the location and spread of invasive species are so costly that they're difficult to undertake at scale.

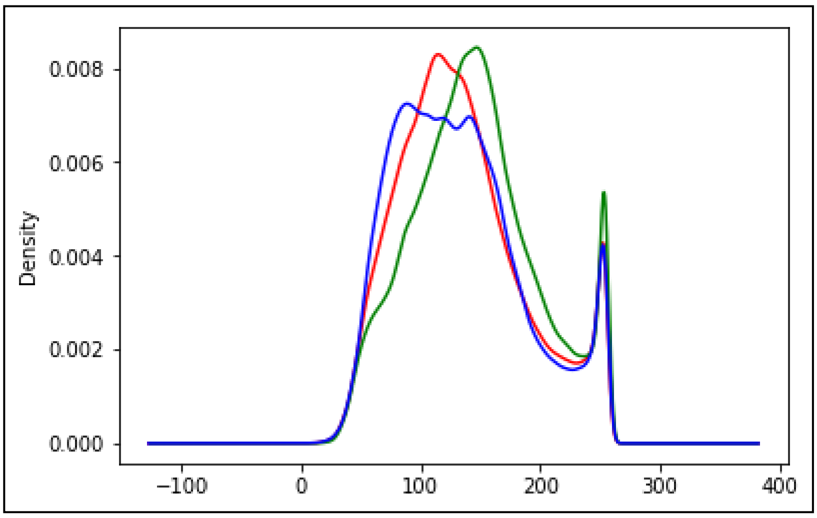
Currently, ecosystem and plant distribution monitoring depends on expert knowledge. Trained scientists visit designated areas and take note of the species inhabiting them. Using such a highly qualified workforce is expensive, time inefficient, and insufficient since humans cannot cover large areas when sampling.

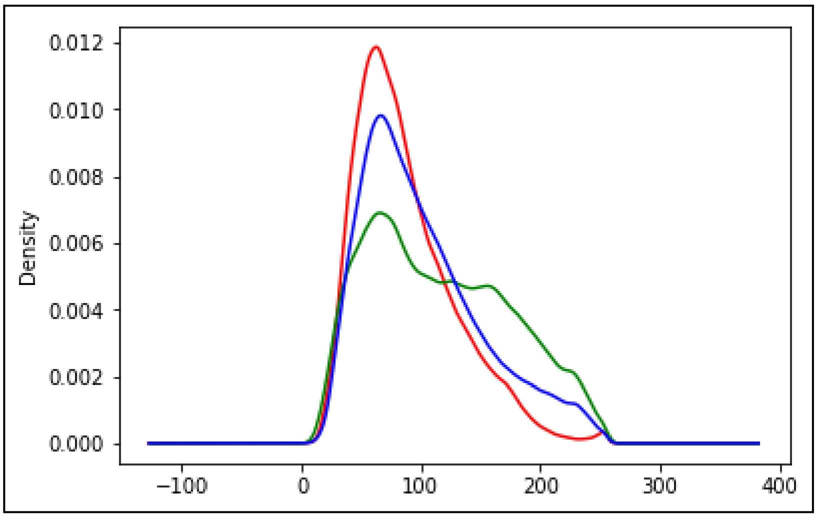
The main consumers for this project will be government agencies around the world interested in preserving natural ecosystems and control the highly invasive plant known as "kudzu". Environmental managers will be able to make more accurate predictions about kudzu invasion and presence, thus making better decisions and increase efficient use of economic resources.

2. Data

The data that i will use for this project comes from a curated dataset provided by kaggle and described as follows:

The data set contains pictures taken in a Brazilian national forest. In some of the pictures there is *Hydrangea*, a beautiful invasive species original of Asia. Based on the training pictures and the labels provided, the participant should predict the presence of the invasive species in the testing set of pictures.





Looking at these plots we can clearly observe that there is a difference in color channel density between an invaded and not invaded site. This first image exploration give us a better idea on the feasibility that a model can distinguish between a forest plot that is free from kudzo and plots where kudzu has invaded.

3. Analysis

Considering the nature of my project and data I will not need to implement inferential statistics as an initial step.

My approach to solve this problem will be based on deep learning algorithms. I will use a subset of the images provided to train the algorithm and test for its accuracy. After some exploration I have decided to use **Convolutional Neural Networks (CNNs)**.

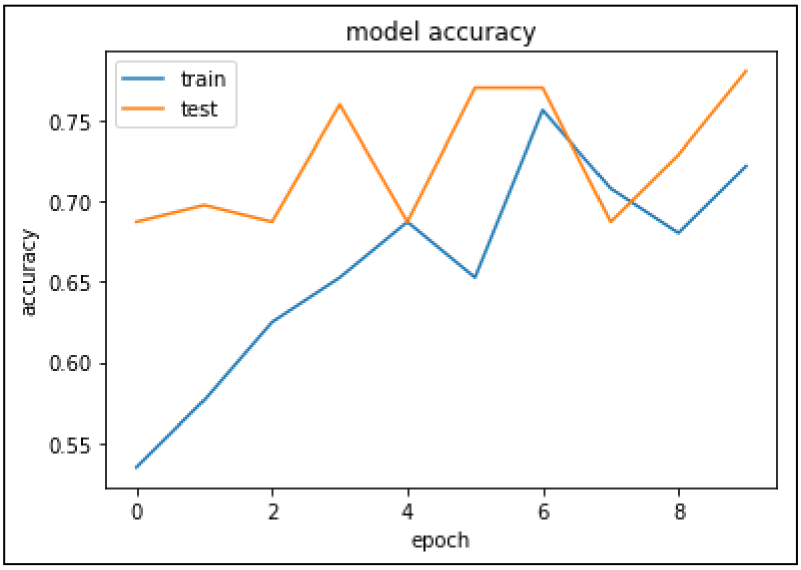
4. Model and Results

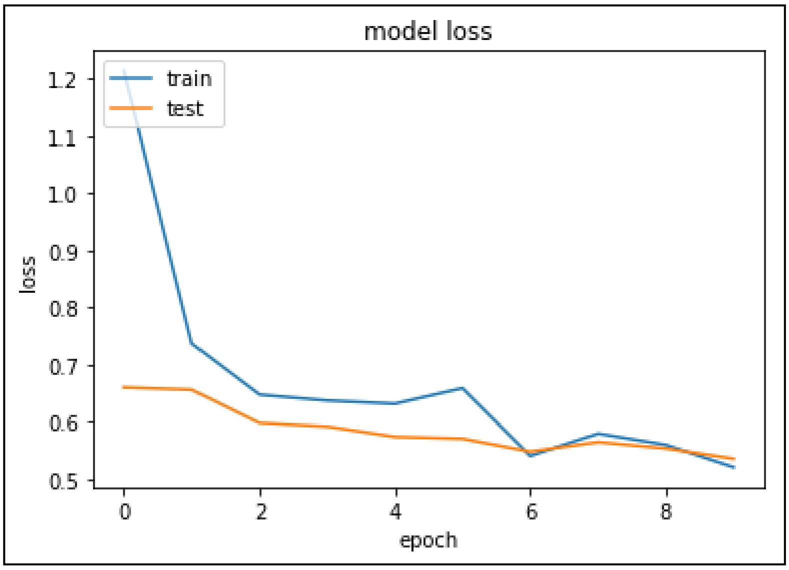
```
Model: "sequential"
_________________________________________________________________
Layer (type)             Output Shape           Param #
=================================================================
conv2d (Conv2D)            (None, 864, 1152, 12)     336
_________________________________________________________________
conv2d_1 (Conv2D)          (None, 862, 1150, 24)     2616
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 431, 575, 24)      0
_________________________________________________________________
conv2d_2 (Conv2D)          (None, 429, 573, 24)      5208
_________________________________________________________________
dropout (Dropout)          (None, 429, 573, 24)      0
_________________________________________________________________
flatten (Flatten)          (None, 5899608)         0
_________________________________________________________________
dense (Dense)              (None, 48)              283181232
_________________________________________________________________
dropout_1 (Dropout)        (None, 48)              0
_________________________________________________________________
preds (Dense)              (None, 1)               49
=================================================================
Total params: 283,189,441
Trainable params: 283,189,441
Non-trainable params: 0
```

```
Train on 144 samples, validate on 96 samples
Epoch 1/10
144/144 [==============================] - 175s 1s/sample - loss: 1.2140 - accuracy:
0.5347 - val_loss: 0.6605 - val_accuracy: 0.6875
Epoch 2/10
144/144 [==============================] - 183s 1s/sample - loss: 0.7371 - accuracy:
0.5764 - val_loss: 0.6565 - val_accuracy: 0.6979
Epoch 3/10
144/144 [==============================] - 170s 1s/sample - loss: 0.6479 - accuracy:
0.6250 - val_loss: 0.5985 - val_accuracy: 0.6875
Epoch 4/10
144/144 [==============================] - 138s 960ms/sample - loss: 0.6377 - accuracy:
0.6528 - val_loss: 0.5919 - val_accuracy: 0.7604
Epoch 5/10
144/144 [==============================] - 153s 1s/sample - loss: 0.6322 - accuracy:
0.6875 - val_loss: 0.5742 - val_accuracy: 0.6875
Epoch 6/10
144/144 [==============================] - 152s 1s/sample - loss: 0.6592 - accuracy:
0.6528 - val_loss: 0.5709 - val_accuracy: 0.7708
Epoch 7/10
144/144 [==============================] - 137s 950ms/sample - loss: 0.5411 - accuracy:
0.7569 - val_loss: 0.5486 - val_accuracy: 0.7708
Epoch 8/10
144/144 [==============================] - 607s 4s/sample - loss: 0.5798 - accuracy:
0.7083 - val_loss: 0.5650 - val_accuracy: 0.6875
Epoch 9/10
144/144 [==============================] - 153s 1s/sample - loss: 0.5603 - accuracy:
0.6806 - val_loss: 0.5541 - val_accuracy: 0.7292
Epoch 10/10
144/144 [==============================] - 113s 782ms/sample - loss: 0.5214 - accuracy:
0.7222 - val_loss: 0.5362 - val_accuracy: 0.7812
```

Test loss: 0.5362207492192587
Test accuracy: 0.78125

Eval loss: 0.5362207492192587
Eval accuracy: 0.78125


model accuracy
train
test
0.75
0.70
0.65
0.60
0.55
accuracy
0
2
4
6
8
epoch


model loss
train
test
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
loss
0
2
4
6
8
epoch

First five probabilities:
[[0.80146617]
 [0.59831256]
 [0.8071442]
 [0.7685373]
 [0.48514572]]

First five class predictions:
[[1]
 [1]
 [1]
 [1]
 [0]]





Although, at first sight the model seems to have failed. The results are highly indicative that a CNN model needs a greater number of samples and epochs to train properly and learn to differentiate among the two types of vegetation for this problem. Given the extraordinary results of CNN models on image recognition I am am certain that the model presented in this project works effectively and will allow conservation managers to make informed decisions for the control of the highly invasive Kudzu.

References:

Buddenhagen CE, Renteria JL, Gardener M, et al (2004) The Control of a Highly Invasive Tree Cinchona pubescens in Galapagos. Weed Technology 18:1194–1202

Kettenring KM, Adams CR (2011) Lessons learned from invasive plant control experiments: a systematic review and meta-analysis: Invasive plant control experiments. Journal of Applied Ecology 48:970–979. https://doi.org/10.1111/j.1365-2664.2011.01979.x

Maguire LA (2004) What can decision analysis do for invasive species management? Risk Analysis 24:859–868

Pyšek P, Richardson DM, Pergl J, et al (2008) Geographical and taxonomic biases in invasion ecology. Trends in Ecology & Evolution 23:237–244. https://doi.org/10.1016/j.tree.2008.02.002