Machine Learning Engineer Nanodegree

Capstone Proposal

Divyansh Sharma

June 6th, 2017

Proposal

Domain Background

Tangles of 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. Hydrangea common names hydrangea or hortensia is a genus of 70–75 species of flowering plants native to southern and eastern Asia (China, Japan, Korea, the Himalayas, and Indonesia) and the Americas. By far the greatest species diversity is in eastern Asia, notably China, Japan, and Korea. Most are shrubs 1 to 3 meters tall, but some are small trees, and others lianas reaching up to 30 m (98 ft) by climbing up trees. They can be either deciduous or evergreen, though the widely cultivated temperate species are all deciduous

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.

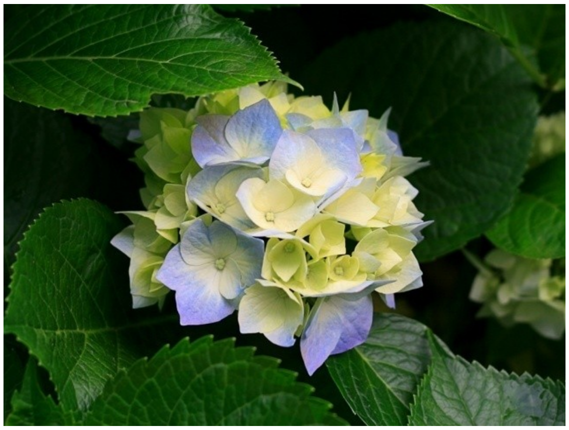

Fig 1. Invasive Hydrangea

There are many government organizations that monitor invasive species like hydrangea and other plants using various methods like mapping and tracking different areas, creating an inventory etc. Also many private organizations like astron are using drones, big data and computer learning to detect and predict invasive species.

This data set of foliage and forest images is publicly available on Kaggle to accurately identify hydrangea in the images. Techniques from computer vision alongside other current technologies like aerial imaging can make invasive species monitoring cheaper, faster, and more reliable.

Many researches in the area of classification in ecology has been done using random forest [(1),]and several other models including factor analysis, climate envelope, machine learning and expert model has been employed in the area of ecological classification with a comprehensive list available at arXiv. [(2)]Also a feature learning based approach [(3)] on high resolution images taken from UAV (unmanned aerial vehicles) is being done in the area of invasive species monitoring.

An extensive research in image classification using deep convolution network is possible through ImageNet, which is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. A deep convolution network [(4)] for image classification using ReLu (Rectifier Linear Unit) and max pooling is discussed.

The aim of this project is to build a convolution neural network that classifies the presence of the invasive hydrangea in the image.

Problem Statement

The foliage and forest images data set was provided by certain individual contributors namely, Christian Requena Mesa, Thore Engel, Amrita Menon , Emma Bradley who are environmentalist and work for monitoring invasive species.

The goal of the project is to identify whether or not the image consist of invasive species hydrangea or not, thus this problem is a binary image classification problem.

The goal is to train a CNN that would be able to classify whether or not invasive species hydrangea is present or not.

Datasets and Inputs

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.

The data set consists:

1. **Train Data**: A collection of 2295 images of size 1144(Height) x 866(Width) pixels, each having three channels for red, green, blue (RGB) in jpeg format.
2. **Train Labels**: A comma separated file (csv) containing name of the pictures (numbers) and class of image corresponding to that number (0 represents Hydrangea not present and 1 represents Hydrangea present)..
3. **Test Data**: A collection of 1531 images of size 1144(Height) x 866(Width) pixels, each having three channels for red, green, blue (RGB) in jpeg format ready to be labeled by the learning algorithm.

Training data consists 1448 images labeled with class 1,i.e. 1448 images contains invasive species and remaining 867 images represents class 0,i.e. invasive species not present. This represents that the classes are not balanced; balanced data is generally good for classification as machine learning algorithms work on majority rule. In scenarios like cancer detection where data can be highly imbalanced like 98% data contains no cancer cases and 2% data contains cancer, there is a need of data balancing and re-sampling. Notice that here the imbalance is not much that can create a bias for a particular class so balancing is not really necessary in this case. However, one case keep track the performance of unbalanced classification by using Precision/Recall which we will be using by creating a confusion matrix for the classifier built. Also, as Kaggle's testing set does not contain labels, we will be using 10 fold cross validation over the training dataset.

Solution Statement

As deep learning is very effective over the years in image classification a convolution neural network will be used to predict the required class label. We will be using transfer learning technique which is nothing but using weights from pre-trained neural network over large data sets. Many such CNN's like VCG16, VCG19,Inception-v3, Xception , ResNet50 etc. trained over imagenet challenge are available for public use .As CNN's require more data while training and as available data set is limited we will be using image augmentation techniques like rotating ,flipping ,Gaussian blur ,Inverting colors, cropping etc. on training data set images which are easily available through image processing libraries like opencv to generate few more images to train our CNN and fine tune its efficiency. As training pre-trained CNN like VCG19 etc are computationally costly we will be resizing and centering the image and working on grey scale images to make training computationally cheaper and feasible.

Benchmark Model

K-Nearest Neighbors: A k-nearest neighbor model was trained on the color histogram model of the training dataset with Euclidian distance as the distance metric. Different values of k (Nearest Neighbors) were experimented and a graph between accuracy and k was plotted to check how accuracies were changing when changing k. After evaluating this optimum k from some set of selected k, 5 was chosen. The above model with selected k was used to predict against testing data and a score of 0.64540 was obtained on the kaggle platform.

A well designed neural net would be able to beat both naïve benchmarks easily and also k-nearest neighbors considering that even KNN surpasses both the naïve benchmarks easily.

Evaluation Metrics

The evaluation metric used by Kaggle to score in this competition is Area under ROC curve also called as **receiver operating characteristic curve**. ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

$$TPR=\frac{TP}{TP+FN} \qquad FPR=\frac{FP}{FP+TN}$$

Submitted values will be evaluated on area under ROC curve between the predicted value and the observed target. To combine the FPR and the TPR into one single metric, we first compute the two former metrics with many different instances, then plot them on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called ROC curve, and the metric we consider is the AUC of this curve, which we call AUROC.

Project Design

- Programming Language : Python 2.7
- Libraries : Keras, Opencv , Tensorflow , Scikit-learn
- Workflow :
  - Establishing baselines with above stated benchmarks i.e. K-Nearest Neighbors for comparison.

- Preprocessing images by converting them to grayscale, centering them and resizing for decreasing computational costs.
- Training a small neural network from scratch for further comparison with transfer learning and pre trained networks.
- Extracting features from pre trained network by removing last fully connected layer and treating the remaining layers as fixed feature extractor for the dataset then finally training a linear SVM or softmax classifier on those extracted feature for comparison. Since the data is small, it is likely best to only train a linear classifier. Since the dataset is very different, it might not be best to train the classifier form the top of the network, which contains more dataset-specific features. Instead, it might work better to train the SVM classifier from activations somewhere earlier in the network. (5)
- Fine tuning convolution network by tuning the weights of the pre trained network by continuing the backpropagation. It is possible to fine-tune all the layers of the CNN, or it's possible to keep some of the earlier layers fixed (due to over fitting concerns) and only fine-tune some higher-level portion of the network.
- Optionally, trying different pre trained networks like ResNet, Xception etc and fine tuning them.

References

1. *RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY.* **D. Richard Cutler, Thomas C. Edwards Jr.,Karen H. Beard,Adele Cutler,Kyle T. Hess,Jacob Gibson,Joshua J. Lawler.** 1 November 2007, ECOLOGY | Ecological Society of America.

2. **Jane Elith(School of BioSciences, The University of Melbourne 3010.Australia).** Predicting distributions of invasive species. [Online] 2011. https://arxiv.org/ftp/arxiv/papers/1312/1312.0851.pdf.

3. *Feature Learning Based Approach for Weed Classification Using High Resolution Aerial Images from a Digital Camera Mounted on a UAV.* **Calvin Hung, Zhe Xu and Salah Sukkarieh.** Sydney : remote sensing, 2014, Vols. Remote Sens. 2014, 6, 12037-12054; doi:10.3390/rs61212037. ISSN 2072-4292.

4. *ImageNet Classification with Deep Convolutional.* **Alex Krizhevsky, Ilya Sutskever,Geoffrey E. Hinton.** s.l. : Conference on Neural Information Processing Systems.

5. **Karpathy, Andrej.** CS231n Concolutional Neural Networks for Visual Recognition. [Online] http://cs231n.github.io/transfer-learning.