Iliana Maifeld-Carucci, Alexa Giftopoulos, Bryan Egan
Machine Learning II
Final Project Proposal

Problem

What problem did you select and why did you select it?

Plankton are a necessary component of our ecosystem. They are small and microscopic organisms that float or drift in the ocean, providing a critical food source for larger organisms as they are the first link in the aquatic food chain. Since they are the primary food producers for ocean life, a decrease or loss of these fundamental species could detrimentally impact our ecosystem. Thus, it is not only crucial to measure and monitor plankton population with underwater cameras, but to be able to classify the high volume of images rapidly, as manual analysis would take about a year for images captured in one day. Therefore, to address this efficiency problem regarding image recognition, we will be applying deep learning techniques to build a convolutional neural network to classify images of various plankton in a timely and accurate manner.

Data

What database/dataset will you use? Is it large enough to train a deep network?

We will use data that was compiled by Oregon State University's Hatfield Marine Science Center. They compiled nearly 50 million plankton images over an 18-day period back in 2014. They then prepared a collection of nearly 30,000 images complete with labels for a training set. These 30,000 images will be split into a 70/20/10 Train-Validation-Test Split to run and test the model.

https://www.kaggle.com/c/datasciencebowl/data

Network

What deep network will you use? Will it be a standard form of the network, or will you have to customize it?

We will use a standard Convolutional Neural Network. This is because we have multi-channeled images that we are classifying so CNNs will help to reduce the number of parameters and make our computation more efficient.

Framework

What framework will you use to implement the network? Why?

We will be using Keras to implement our CNN network. Keras is reliable and flexible, as it can be deployed on a greater range of platforms than any other network. Additionally, it has a strong multi-GPU support and distributed training support, which is ideal for handling complex networks such as CNN.

Reference Materials

What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

The initial write-up includes a large portion of background for the study as well as a PDF file of the species and subspecies breakdown so we can correctly classify each image.

https://keras.io/

Keras has a substantial and robust documentation that we will use to help us create the model. It includes how to correctly standardize data, how to add additional elements such as enhancing and whitening, as well as centering and normalizing images.

Metrics

How will you judge the performance of the network? What metrics will you use?

Overall, we will be analyzing the predictive accuracy of our model while we manipulate the number of layers, neurons, and batch size in our network to determine the ideal balance in providing us accurate and timely results.

We will analyze the performance of the network using the ROC curve and confusion matrix to see how the overall numbers compared to their predicted values. We will use stochastic gradient for optimization.

Schedule

Provide a rough schedule for completing the project.

Project Proposal- October 31st
Nov. 5th- Nov 26th: Fully understand data structure, sizes, and Keras background
Nov 27th - Dec 4th: Model building and implementation. Making sure the code is functional.
Dec 4th - Dec 10th: Model tweaking and parameter tuning. Finalizing report write up and presentation

Presentation- December 10th

Report- December 10th