Predicting ocean health, one plankton at a time

Peeyush Agarwal
peeyusha@iitk.ac.in
12475

Abhilash Kumar
abhilak@iitk.ac.in
12014

March 16, 2015

1 Motivation

Plankton are critically important to our ecosystem. Aside from representing the bottom few levels of a food chain that supports commercially important fisheries, plankton ecosystems play an important role in the bio geochemical cycles of many important chemical elements, including the ocean's carbon cycle. Loss of plankton populations could result in ecological upheaval as well as negative societal impacts, particularly in indigenous cultures and the developing world. Plankton's global significance makes their population levels an ideal measure of the health of the world's oceans and ecosystems.

Traditional methods for measuring and monitoring plankton populations are time consuming and cannot scale to the granularity or scope necessary for large-scale studies. A better approach is to use underwater imagery sensors for capturing microscopic, high-resolution images over large study areas. These images can then be analyzed to assess species populations and distributions.

Manual analysis of the imagery is infeasible as it could take a year or more to manually analyze the imagery volume captured in a single day. Automated image classification using machine learning tools will allow analysis at speeds and scales previously thought impossible.

2 Objective

The objective is to build an algorithm to automate the plankton image identification process. More specifically, the task is to create an algorithm that given an image, assigns class probabilities for various plankton classes.

3 Dataset

For the National Data Science Bowl competition [1], scientists at the Hatfield Marine Science Center have prepared a large collection of labeled images, approximately 30,000 of which are provided as a training set. A test test of approximately 130,000 images is also provided for the competition. This dataset has been used for the project.

4 Challenge

Several characteristics of this problem make this classification difficult:

1. There are many different species, ranging from the smallest single-celled protists to copepods, larval fish, and larger jellies.

2. Representatives from each taxon can have any orientation within 3-D space.

3. The ocean is replete with detritus (often decomposing plant or animal matter that scientists like to call "whale snot") and fecal pellets that have no taxonomic identification but are important in other marine processes.

4. Some images are so noisy or ambiguous that experts have a difficult time labeling them. Some amount of noise in the ground truth is thus inevitable.

5. The presence of "unknown" classes require models to handle the special cases of unidentifiable objects.

5 Methodology

Deep learning approaches have shown good results for computer vision problems [2]. Deep Convolutional Neural Networks [3] were especially successful in object recognition on the ImageNet dataset [4]. We plan to use CNNs for plankton image classification. We also plan to design and optimize our network layers and mix CNNs with other models based on domain and dataset. Since the approach is computationally expensive and almost infeasible without GPUs, we also plan to compare it against some computationally less expensive approaches like random forest.

References

[1] *Data Science Bowl Challenge*
`http://www.datasciencebowl.com/`
`www.kaggle.com/c/datasciencebowl`

[2] Ciresan, Meier, Masci, Gambardella, Schmidhuber *Flexible, High Performance Convolutional Neural Networks for Image Classification.* IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. No. 1. 2011.

[3] Lecun Y. , Bottou L. , Bengio Y. , Haffner P. *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 86(11),2278 - 2324,1998

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems. 2012.