Machine Learning Applications for International Ocean Discovery Program Geoscience Research

Peter Flaming1, Brandon De La Houssaye1, Quinton Nixon1, and Gary Acton2

1 Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {pflaming,bdelahoussaye,qnixon}@smu.edu2 International Ocean Discovery Program, Texas A&M University, College Station, Texas 77845 USA acton@iodp.tamu.edu http://iodp.tamu.edu/staﬀdir/indiv/acton/

**Abstract.** In this paper, we present the use of a machine learning in order analyze high-resolution still images gathered from decades of scientific ocean drilling data. The International Ocean Discovery Program (“IODP”) – which is part of the National Science Foundation (“NSF”) – has been overseeing the JOIDES Resolution scientific drillship as it travels the oceans and takes samples of the earth and rock beneath the ocean floor. While the images have been in existence for some time, the means to extract the colors in the image into a usable data file for scientists has thus far eluded the IODP and NSF. These sample are what we apply computer vision techniques to in order to classify the colors (based on location in the sample) and then flatten the results into a data format. *Or, put another way, in this paper we present a means of using machine learning in order to transform the color from the sample into useful data so that it may be analyzed by scientists in order to study the history of the earth.* A computer vision model was utilized in order to first develop a Deep Neural Network (“DNN”) to mine and extract important features from the sample photos, and a Recursive Neural Network (“RNN”) was then utilized against the preprocessed DNN data. [NOTE: additional modes & techniques are being explored such as variants of CNN.] The overall goal of the model is to take an image and transform it into a flat data file indicating the shades of color (on a grey scale) against the location in the sample. The conclusion presented is the model built and the indicators of accuracy.

Keywords: IODP · JOIDES Resolution · SST · Machine Learning · Deep Learning · DNN · RNN.

**1 Introduction**

There is significant work undertaken by a broad array of scientists, industrialists, and officials – really, all of humanity – seeking to understand our Earth. A particular hot topic today is the Earth’s climate. Geologists have the ability to explain the present and the past of the Earth by analyzing the dirt, rocks, etc. right under foot. However, as a can be imagined, the sheer volume of information that comes from the samples needs to be a format where the geologists can perform their scientific studies.

As it relates to this paper, of particular interest are samples taken the beneath ocean floor. Historically, the data collected from the samples were able to be integrated parametrically in order to describe the present climate. However, with new technologies capturing data in near real-time and in high resolution it presents nonparametric relationships that make it hard for users to model the past and future of the Earth’s climate with traditional statistical models.

Lacking in these traditional methods is the ability to expand these relationships nonparametrically to other locations, such as onshore basins, oﬀshore basins, diﬀerent wells, and the emergence of fracking. Thus, as the scientific community has expanded their ability to gather information in real time and with better tools (such as high-resolution photography) a new problem has emerged around how to transform this gathered information into more computationally powered modeling techniques.

As more cores are gathered, and the scientific community seeks to gain better insights into the whole of what the core can tell us by using better information gathering tools, there is a real need to transform this additional information into a format where scientific understanding can occur.

The above summary is so broad encompassing and the problem statement is so high level, that it may feel of hubris. While the presenters of this paper maintain awareness of such, they also do not limit the potential contributions contained herein towards those broader goals. Nevertheless, in a practical sense, this paper can be described as containing a ‘proof of concept’ computer vision model that focuses on transforming one element of the information to be gleaned from the sample (the high-resolution photographs) into a data-driven usable format where existing (or soon to be developed) statistical models can be applied in order to better understand the Earth’s history.

In building the model, the key focus is one of scalability. Scientists can – and have – gone through the sample photos and translated the colors into a flat (data) file based on color spectrum using spectrophotometry. Basically, this means that for the core sample, at a given location in the photo, on a grey scale from 0 to 255, what is the color shade? The results are then mapped and plotted, and the resulting data can be statistically analyzed. So, the question at hand is why would a team of data scientists be needed? The answer is that it has been done ‘by hand’ over a painstakingly long and expensive process. What data science can do is develop a model that for a given image, the resulting classification of the image (against a greyscale) becomes automated. The model, once it has been trained to achieve satisfactory accuracy scores, is scalable. That is, the model can be employed in real time to additional core photographs. The end result is a data tidal wave to the population of geologists patiently waiting to their experiments.

In order to develop the computer vision model, we set out to first understand the core sample photographs by deploying DNN techniques along with a RNN technique against the features extracted from the DNN technique. Simply put, we have to first collect the photophraphs and place them in a format whereby the pixels can be analyzed, recorded, and classified accordingly. The team employs a Convolutional Neural Network (“CNN”) for analyzing the information and building a model. A key point worked through as part of this project, was the notion of ‘time-series’. The deeper a core sample goes, the longer that time has passed. i.e., earlier period’s in the earth’s history are represented as lower in the vertical picture. However, there is not a clear indicator from the samples when one time period ends and another begins. For that reason, RNN would be potentially ideal for developing the model, but CNN is also explored given its reliance on fixed-size inputs and outputs (when analyzing pixelated images and their location within the larger photograph).

The end result – the output of the model – is a ‘flattened’ data file that shows on a color scale against a location allowing the possible graphing of the data. The conclusions that can be reached is that a proof of concept computer vision model was developed whereby the images used to train and test the model were correctly assigned XX amount of the time.

The remainder of this paper includes an overview of the data gathered and utilized; the methods and experiments performed; the results (of the model developed); an analysis of those results (i.e., accuracy indications); ethical considerations; and overall conclusions. Finally, this paper contains a summary of potential future work as well as the associated references and appendices.

**3 Dataset and Data Exploration**

The team has received copies of core images for analysis. Upon request, the team will receive additional copies of the images. The dataset needed to build the model are provided by the IODP (Texas A&M) database for use. The images are high-resolution.

The images represent the cross section of the sample taken beneath the ocean floor surface. As the exploration drilling occurs, the sample is pulled up and analyzed as well as cut in half and then photographed. Each image is stored and tagged with the sample location as well as time it was taken (from the Earth).

**4 Methodology and Experiments**

Note: this section is left intentionally blank as not methods have been finalized and experiments have yet to be conducted.

**5 Results**

This section should show the test results when taking the model through the selected test image(s). As no methods or experiments are ready for provision, no results are available as well.

**6. Analysis**

Note: needed is a means to consider accuracy scores as well as any hypothesis testing.

**7. Ethics**

This work can be seen as contributive to larger scientific endeavors focused on better understanding Earth in terms of a variety of considerations including climate. The information, data, and assistance required to perform this work cannot be seen as infringing upon another entity’s labors or proprietary, valuable information. In a very general sense, this work has extraordinarily limited ethical concerns.

**8. Conclusions**

The overall conclusions reached are that the

**9. Future Work**

The model contained in this paper was built and trained using a limited number of sample photographs. As additional core samples are gathered and analyzed, the model should continue to be trained for increased performance in terms of accuracy.

Additionally, as this paper sets out, the model being built is one geared towards computer vision with a focus on transforming image information into data files. Once this is accomplished, and a model is put into production, the possibility for future work is so broad that this paper cannot possibly list in completeness. However, some immediate steps would be possibly deploying the computer vision model to other images sets (for retraining). Additionally, there will be multiple opportunities to continue deploying machine learning methods for other issues facing the geologists (now that they have this data) including: models to infer the contents of the Earth between the location of the samples; predicting characteristics of the ocean the Earth from the color shades of the core; etc.

**10. References and Appendices**

As the analysis progresses, this section will grow exponentially…