UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI
UNDERGRADUATE UNIVERSITY

Group Project

Plankton Classification Platform for Supporting Research Activities of Environmental Scientists

Report 2

Submitted by

LE Nguyen Khoi	USTHBI7-086
NGUYEN Xuan Bach	USTHBI7-021
LE Minh Anh	USTHBI7-007
DOAN Thanh Huy	USTHBI7-084
PHAM Minh Duc	USTHBI7-041

Under the guidance of:

Dr. NGHIEM Thi Phuong ICT Lab

Hanoi, January 2019

Table of Contents

I. Introduction

Plankton is a diverse group of organisms that live in the oceans and cannot swim against the water current [1]. And they have an important role to ecosystem and human life that we cannot undeniable. The carbon cycle of the Earth is almost controlled by these tiny organisms. In order to study plankton's population that is imbalance nowadays because of the ecosystem changed. Underwater camera systems are used to capture microscopic, high–resolution images of plankton over large study areas. However, there are a huge amount of plankton images, it's raises a key question that how we can analyze the collected plankton images in order to assess the population and distribution of plankton species. A natural approach to this question is to manually classify collected plankton images for further human analysis This approach, however, is infeasible since it would be extremely time and cost consuming to manually classify millions of images [2].

Figure 1. an Image of Plankton

1. Context of Project

To overcome this problem, the project has two main goals:

(1) To develop a classification platform to automatically put new plankton images in their correct categories;

(2) To build a plankton website so that the environmental scientists can contribute plankton data for easier use of plankton images in research and education.

2. Main Question

When we were doing our project, there are a lot of questions that come up to us. However, there is a question that our team care most is “Is our website can automatically detect new kind of plankton and tag it to a group in gallery when people upload to website’s data?”

This question come up when we very worry about what if the website can only detect the image from a type of plankton that we had put in the train model, when we are doing the project. Will it detect the new plankton or they will they the new plankton is just some kind of random image come up. We still thinking a about that problem for the time being. And not yet find the answer.

3. Aims and Objectives

- **Aims:**
  - This project is about making a website using knowledge that we learned in the university.
  - To develop teamwork skill with the use of tools like Github and Google Doc.
  - To learn process of making a project.
  - Build a groundwork that can be reused and expanded in the future.
  - Understand the basic background about machine learning and image classification.
  - Successfully build and develop a part of website that image can be upload to website data, and website can detect that image belong to which type of plankton.

- **Objectives:**
  - Apply image classification models for the case of big datasets.
  - Apply image classification models to build plankton classification platform.

- Develop web user interface for environmentalist to use plankton classification platform and let people also can contribute plankton data for easier use of plankton images in research and study

II. Literature Review

Nowadays, there are many kinds of methods in the literature to classify plankton around the internet. One of them that we cannot discuss about is an automated platform for classifying and monitoring phytoplankton that using scanning flow cytometer. This cytometer was designed to analyze the natural properties of plankton. For example, the cytometer is capable of analyzing from small to large planktonic particles (from 1 to 700 µm in diameter and a few mm in length) and relatively large water volumes. The platform manipulates and visualizes data using R programming language. The datasets also include original sampled volume, date, time, and depth at which particles were taken.

This platform was proposed by pomati et al, 2011 [3]. With the idea of a platform that it is able to extract very detail information of planktonic particles for further analysis is very interesting. However, if students or researchers want to have a large amount of plankton datasets for large-scale studies, the platform shows the limit in gaining high efficiency to classify the plankton. This is because the platform has not yet taken into account the advanced data analysis techniques to handle big datasets. So to solve the problem difficult use of plankton images for research and education, we come up with an idea that we create a website, an online plankton analysis system (PAS) for environmentalist to use plankton classification platform. Basically, the system is a web application allowing marine biologists to classify their plankton images using a web user interface. PAS provides the functions for experts to upload their images, process them, extract features, hand-label image, train classifiers, and use those classifiers to automatically label new images [4]. So that people can contribute plankton data in order to help other people want to research and study. And to make thing much easier than pomati et al, 2011 [3] platform by we apply the modern machine learning methods to our website, namely convolutional neural network, in order to gain high performance computing when analyzing large datasets.

III. Scientific Methods

From what we mentioned above that our motivation in this work is to build a classification platform to put new plankton images into their correct categories. In the following, we will present in detail the machine learning methods we use to perform this motivation.

1. Machine learning and image classification

According to Michalski et al, 2013 [5] machine learning is a field of study that provides computers with ability to learn from experience. To the detail, that we talked about that a computer program is said to learn from experience if it can teach itself to grow and change having new data. With the detail, we can tell that machine learning focuses on the development of computer algorithms for transforming data into appropriate action. For example, that make it easier to understand how the machine learning environment work. In the environment, there are available data, statistical models and computing power are simultaneously interacted with each other. As a result, from that, growth in data causes the need of additional computing power, which in turn spurs the development of statistical method for analyzing large datasets. This make it become a loop cycle of advancement allowing even larger and more interesting data to be collected.

Image classification refers to the task of classifying images into appropriate classes or categories that they belong to. In other word, we can say that image classification algorithms perform the press of putting unlabeled images into their class, type or categories that they belong to.

In the following, we will present our study on the machine learning models - convolutional neural network as the image classification models.

2. Image classification models - Convolutional Neural Network.

Convolutional Neural Networks (CNN) is a variant of multilayer perceptron (MLP) introduced by LeCun et al, 1998 [6] to overcome the weakness of MLP. To the detail, like we had mentioned above, it is wasteful to connect each neuron in a hidden layer to only a local region of the input layer. This connection is executed by using a convolution filter over the input layer and the hyper-parameter of such region is called the receptive field of the neuron. The convolution filters act as local filters over the input space and are well-suited to exploit the strong spatially local correlation in the natural images.

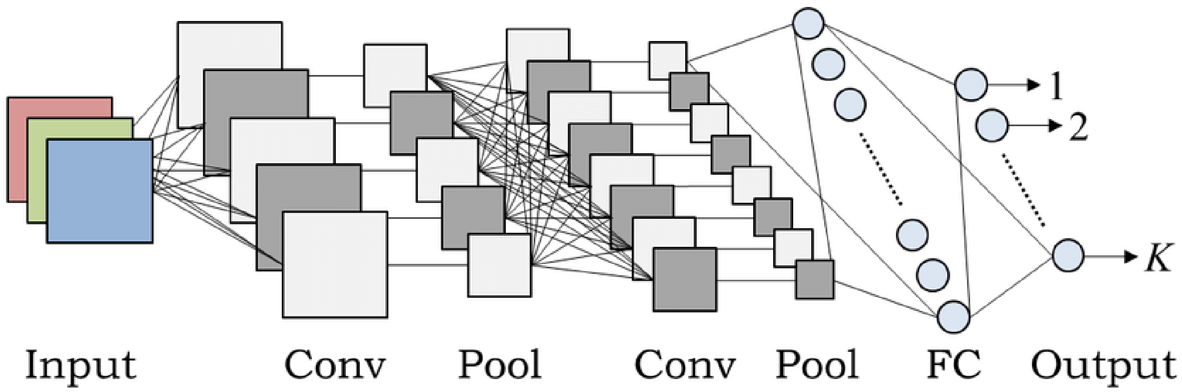

Figure 2. An example of CNN architecture.

3. Web user interface

Since our primary aimed like we mentioned on the title plankton classification platform for supporting research activities of environmental scientists, it is more convenient to provide them with a graphical user interface to use our classification platform. For this, we develop a website acting as a web user interface for our plankton classification platform.

A detail of our design and implementation is presented in the following.

In order to build our front-end for users to interact directly, we use the skill that we had studied from the course web application. We use HTML, CSS, Javascript, Jquery to design our web framework. However, in CSS part, mostly we use high level of CSS technique, namely SASS. We choose SASS for our most thing in CSS because SASS can provide us variable to set and use when we making website interface. Moreover, SASS can help to design web complicated interface.

IV. Experimental Results and Discussion

1. Experimental setup

a. Dataset

Our data presented here are subset of a larger plankton imagery data set contain millions of images of microscopic marine plankton, organized according to category labels provided by researcher at the Wood Hole Oceanographic Institution (WHOI). The images are currently placed into 6 Domains of plankton: Diatoms, Flagellates, Dinoflagellates, Ciliates, Coccolithophore, Miscellaneous. To the detail, in each domain, there are many types of plankton. There are total 103 types of plankton in 103 folders. These images were collected in situ by automated submersible imaging-in-flow cytometry with an instrument called imaging FlowCytobot (IFCB) [7]. Our data that we use on this project take from 2010 to 2014 (There are total 1,990,000 images in 103 folder)

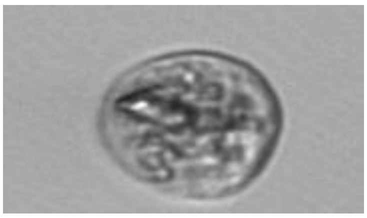

Figure 3. An image of plankton in our data set

In our plankton data set that we extracted from the raw data were sorted into 103 plankton classes, so that we split 6:3:1 into train, validation and test datasets.

b. Data Preprocessing

In our dataset, there are many images with different resolution, the smallest ones are less than 50 x 50 pixels, whereas the largest ones are up to 1000 x 1000 pixels, even some images will come up with 500 x 1000 pixels or 148 x 200 pixels. Therefore, we decided that we rescaled in fixed dimension: 128 x 128 pixels in order to unified all images on a resolution. To improve the stability and convergences of the model.

However, there are some folder of plankton datasets only contain few images or less than 200 images. With this problem, this will cause us overfitting our training model. So, we used data augmentations to handle randomly during training for preventing overfitting. The parameters of data augmentations are presented below:

- Rotation: random with angle between 0 – 180 degree
- Width shift and height shift: randomly translate pictures 0.2 vertically or 0.2 horizontally
- Rescale: Target values between 0 and 1 by scaling with a 1/255
- Shear range: randomly applying 0.2 shearing transformations [8]
- Zoom range: 0.2 for randomly zooming inside pictures
- Horizontal flip: randomly flipping half of the images horizontally - relevant when there are no assumptions of horizontal asymmetry. We set to True.
- Fill mode: it’s a strategy used for filling in newly created pixels, which can appear after a rotation or width/height shift. We set nearest.

c. Training and Validating

Like we just mentioned above we divided our data set with the ratio is 6:3:1. The training process with bath size is 8, and samples that we put to train per epoch is 1000 and validation steps is 500. The learning rate is initialized at 0.0001 and a learning rate decay of 0.0001 is applied after each epoch.

2. Result and discussion

Effect of number of neural network layers:

Overall, the accuracy of all neural network training until now have passed 90% (see appendix) and the time of training also increase with the number of layers as a function like n^a. We have trained the model for many times. The first few times, the train accuracy is very high and loss is below 20%. However, our validation accuracy is not high as we expect, it’s only fluctuations in the average threshold, about 70% and the validation loss always equal to validation accuracy. Then when we use computer to predicted it, it is not show us the high result, there are too much images that computer can not detected. We took a domain to train for out example, in the end our data from that model has an overfitting. However, we have come up with a solution and still working on it right now.

Plankton characteristics classified by the platform:

As the result is scientific name of plankton, more characteristics of that plankton can be known. According to Jane M. Browles [9] and Jerry G. Johnson [10], the scientific name is in Latin name and consists of two parts: the genus or generic name and the specific epithet. So that, from the name of plankton, the properties ò plankton can be almost understood and this is very useful for further research.

Website interaction:

Our website, basically is a hybrid website, that combine a news website about plankton and a huge gallery that people can access to it, and also people can contribute our data through our website. To the detail, with the part news website, people will just use it for update their information about new story, or new discovery about new kind of plankton or something else that relate to plankton. And about the gallery part, people will access to it to find what their plankton images that they want to see, also, people can upload their image about plankton to this website with only few easy steps.

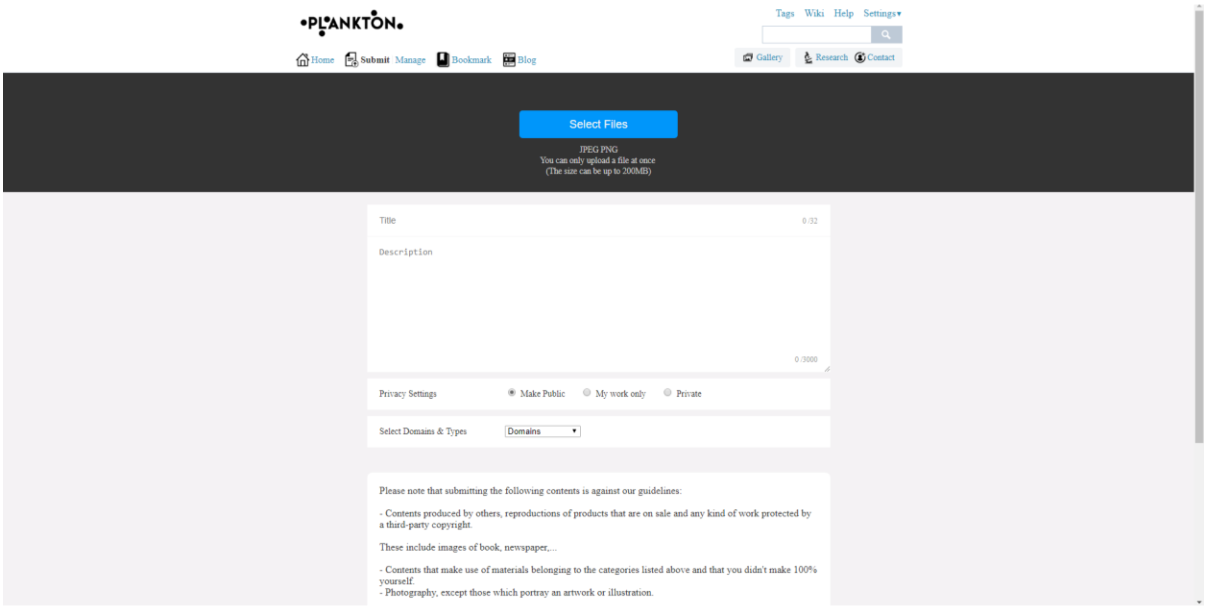

Figure 4. Web interface

If the image that already that exist in the website database, the website will immediately warn the user that this image had already had on our data, so it can be uploaded to our data.

Also base on "intellectual property", we applied some law to our website to write term of use. In order let people understand that people cannot reup images that already exist in our database.

V. Conclusion and Future Works

In the project, the accuracy of the experiment that until now, we have is over 90%. However, that only the number that we use for only a domain in our data. We haven't yet used all the domain in WHOI's plankton dataset for training. We are trying to make it work for real at the moment. But to comparing with the classification rate achieved by machine learning method to human accuracy that over 90% is a fine number.

For the future work, this would be possible to improve the activities of the website.

For the time being, we only make it can be uploaded the image to our website and move it to our database so people can used it. In the future, we also want that people from outside also can write article and upload it to blog page in our website. So people can learn something from the point of view of the people write it and uploaded to our website.

VI. Management of the project at the end

Activities	Plan	Expected results	Done
Research of Plankton	+) Understand overall characteristic and a bit knowledge of plankton +) Understand 6 domains of plankton +) Prepare the content about plankton	+) Understand overall characteristic and a bit knowledge of plankton +) Understand 6 domains of plankton +) Prepare the content about plankton	+) Understand overall characteristic and a bit knowledge of plankton +) Understand 6 domains of plankton +) Prepare the content about plankton
Preparing detailed information about plankton and code	All required software is installed	Everything is ready for testing	+) Collected and checked all the articles and documents +) Downloaded database (103 types of plankton) from Wood Hole Oceanographic Institution (WHOI)
Model Development	+) Data exportation +) Image preprocessing + Apply machine learning for building a model	The model has accepted classification score	The model reaches up to over 90% (For only one domain that we are now testing, not the whole data) – It's a good result, since there are many mistakes that we had fixed.
Website Development	+) Design website of plankton classification	+) Simple and responsive user interface +) Can use pre-built model to recognize	Website is simple and friendly to user

		plankton	
Writing report and preparing for the final presentation	+) Preparing the slides for presentation. +) Practicing the presentation +) Think and ready for every possible question that may come up +) Writing report	+) Completed on time +) Well prepared the slides for presentation +) All important information is covered +) Come up with many questions that may asked during presentation	+) Well checked before submitting report. +) Prepared well for presentation +) All important information is covered +) Be sure all members understand well their own task and of others

References:

[1] Castro Pond, Stephen, and George L. Pickard. Introductory dynamical oceanography. Elsevier, 2013

[2] Culverhouse, Phil F., et al. "Do experts make mistakes? A comparison of human and machine identification of dinoflagellates." Marine Ecology Progress Series 247.17-25 (2003):5.

[3] Pomati, Francesco, et al. "An automated platform for phytoplankton ecology and aquatic ecosystem monitoring. "Environmental science & technology 45.22 (2011): 9658-9665.

[4] Plankton Analysis System, http://vis-www.cs.umass.edu/~pas/, accessed February 2016

[5] Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell eds. Machine learning; An artificial intelligence approach. Springer Science & Business Media, 2013.

[6] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[7] Building an Automated Underwater Microscope. A conversation with biologist Heidi Sosik (http://www.whoi.edu/oceanus/feature/building-an-automated-underwater-microscope)

[8] Shear mapping, from Wikipedia, the free encyclopedia (https://en.wikipedia.org/wiki/Shear_mapping)

[9] Jane M. Browles, the Name of Plants, 1996.

[10] Jerry G. Johnson, The Classification of Living things, 2005.

Appendix:

Result Table

Number of neural network layer	Time to train (hr)	Accuracy (%)	Loss (%)
1	6	70	70
2	About 7	78	72
3	8	84	44
4	8	89	30
5	About 9	94	29

Result after testing on a domain (The domain we also reduce less images than the original number of original images from our data to train and test it faster)

True Dictyocha3: 16
False Dictyocha3: 4
True Dinobryon5: 19
False Dinobryon5: 1
True Euglena3: 11
False Euglena3: 9
True flagellate_sp33: 19
False flagellate_sp33: 1
True Phaeocystis3: 18
False Phaeocystis3: 2
True Pyramimonas_longicauda3: 11
False Pyramimonas_longicauda3: 9

Hint:

- **True** mean the data images that we test that computer recognized it and put in the
  right type of plankton dataset
- **False** mean the data images that we test that computer cannot recognized it and cannot put them in the type of plankton that they must belong to