A Deep-Learning pipeline for diatom detection and classification

End of studies project of student FAURE--GIOVAGNOLI Pierre

Supervised by Dr. PRADALIER Cédric (Georgia Tech) and Dr. SOLNON Christine (INSA)

Supported by the UMI 2958 CNRS-GT and UMR 7360 CNRS-UL laboratories in Metz

INTRODUCTION

Diatoms are a type of unicellular microalgae found in all aquatic environments. Their great diversity and ubiquity make these organisms recognized bio-indicators for monitoring the ecological status of watercourses, particularly in the context of the implementation of the European Water Framework Directive.

With this project, we address the two following topics:

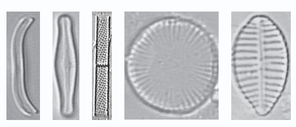
DIATOM DETECTION → **DIATOM CLASSIFICATION**

DIATOM DATASETS

Atlas (2020)

Atlas is the main diatom dataset of this project. The images have been extracted from 3 DREAL diatom atlases gathering samples from the hydrographic basin Rhin/Meuse [1] [2][3]. The main challenge of this process was to extract the right images with their respective labels, some atlas needing extensive segmentation tasks and many filters to reduce manual post-processing.

The Atlas dataset is composed of 157 taxa with a median of 21 images per taxon.

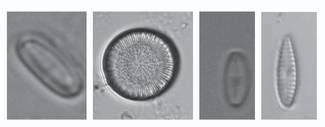


Aqualitas (2017)

In 2017, [4] proposed an update on diatom classification reaching 99.55% of accuracy with the Alex-Net convolutionnal neural network. They achieved those scores with their own dataset created in partnership with the Spanish National Research Council.

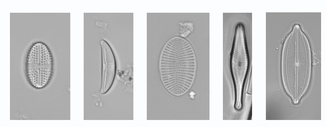
In this project we propose on update on their work by applying the latest CNN advances in image classification on their dataset.

The Aqualitas dataset is composed of 100 taxa with a median of 100 images per taxon.



ADIAC (2002)

The ADIAC project [5] sets the first state of the art reference for automatic diatom classification and made a robust diatom dataset available to the public. The original subsets used for their experiments not being available anymore but a following paper [6] published in 2011 used 3 new subsets composed of 38, 48 55 taxa that we will name respectively ADIAC38, ADIAC48 and ADIAC55.



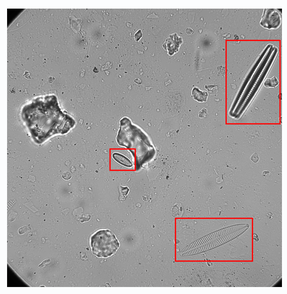
DIATOM DETECTION

Goal

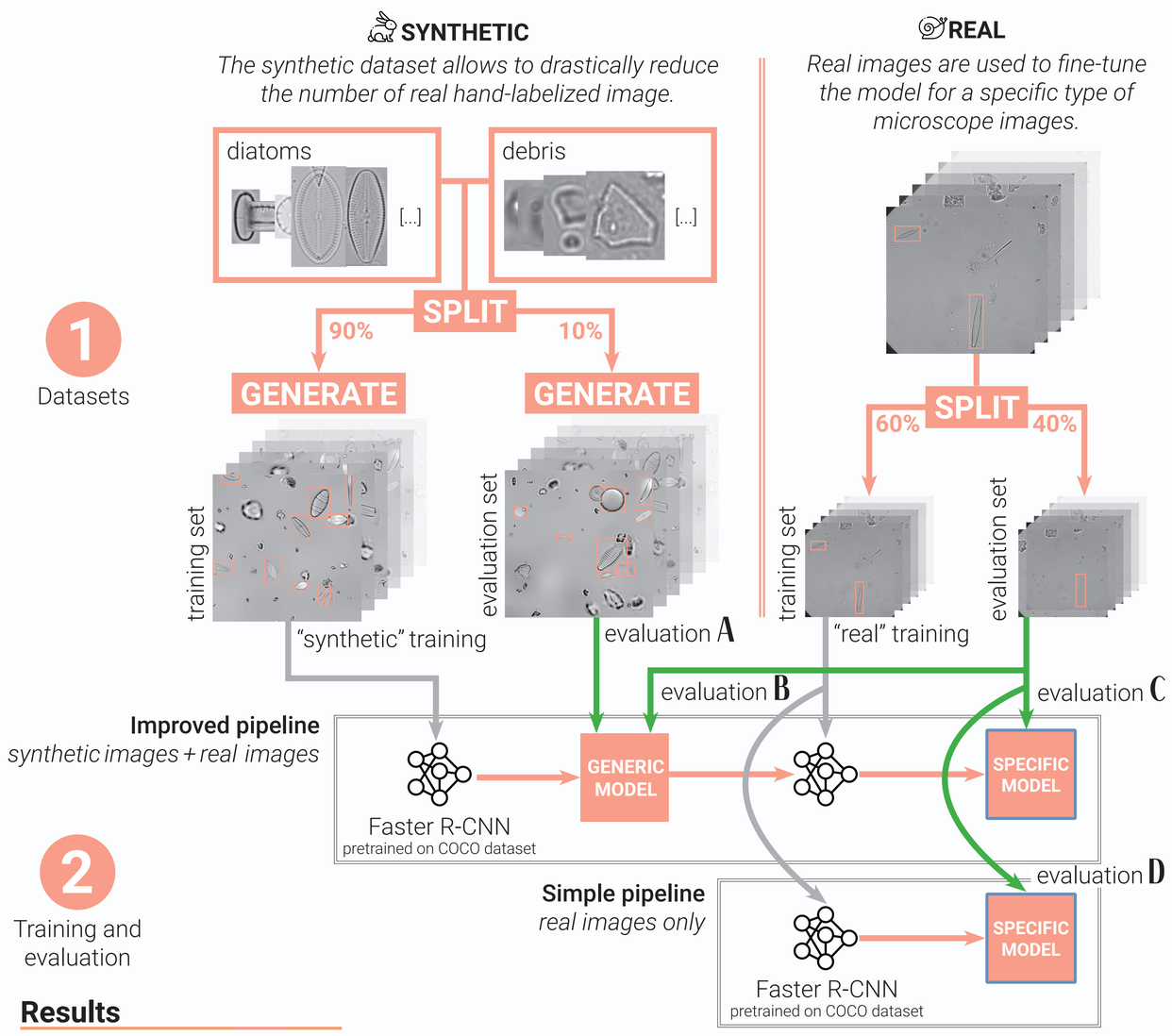
By detection, we understand the **localization** of diatoms on a **microscope image**. Hence, the first objective of this project is to apply a state-of-the-art object detection architechture to detect diatoms in light microscopy images. An example of such image with framed diatoms is visible on the right.

This approach is new for two reasons:

- it uses a **deep learning object detection architecture** for microorganism detection
- the training is made using a dataset of **synthetic** multi-taxa microscope images



Process



Results

	A	B	C	D
Type	synthetic	real	real	real
Images	3000	185	185	185
$AP_{IoU=0.50:0.95}$	0.876	0.247	0.612	0.515
$AP_{IoU>0.50}$	0.990	0.580	0.857	0.768
$AP_{IoU>0.75}$	0.965	0.117	0.737	0.623
$AR_{max=1}$	0.097	0.215	0.333	0.297
$AR_{max=10}$	0.871	0.397	0.724	0.652
$AR_{max=100}$	0.905	0.426	0.728	0.667

For evaluation, we used COCO's set of metrics as they are widely spread and cover a wide range of criteria.

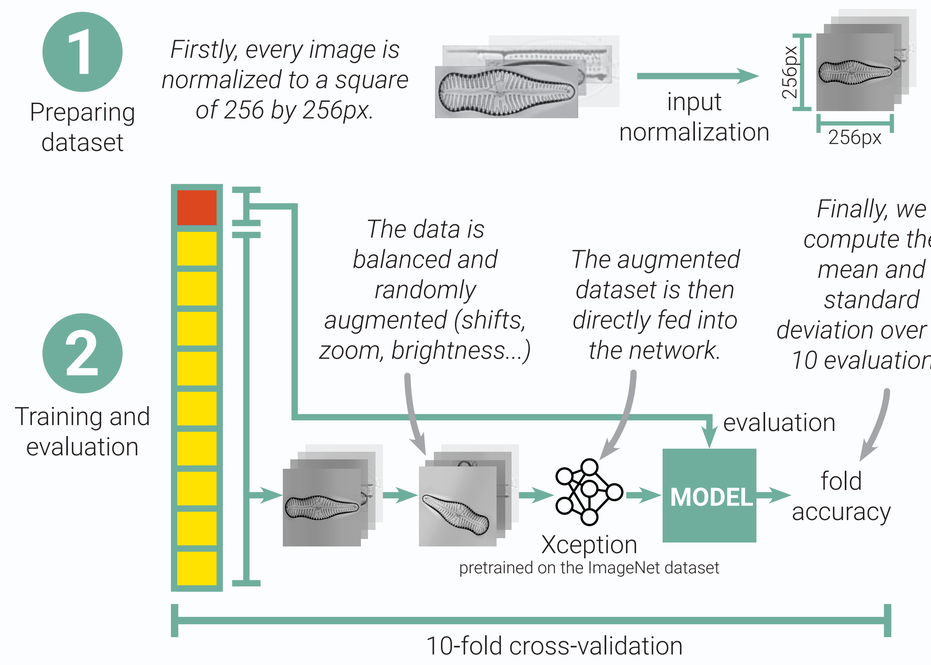
Seeing results of evaluation B, It is clear for now that the generic model can not be applied out of the box to any type of microscope image and that the fine-tuning process on real images is essential. However, thanks to the synthetic dataset, we have a significant gain with with **C** in comparison to **D**! It means that the synthetic dataset allows to better generalize and spare tedious manual labelization.

DIATOM CLASSIFICATION

Goal

Thousands of diatom taxa have been discovered to this day and identifying them is of great interest for biologists as they reveal a lot on their environment. Manual diatom classification is a difficult and time-consuming task and a lot of studies worked on automating the process. In this study, we propose an update on the subject using a state of the art CNN image classifier (Xception) allowing to extract high level image features.

Process



Results

	Atlas	Aqualitas	ADIAC55	ADIAC48	ADIAC38	
#taxa	166	80	55	48	38	
Median #images/taxon	51	94	20	20	21	
Accuracy	0.9265	0.9362	0.9672	0.9735	0.9713	10-fcv
Previous best accuracy	Ø	0.9951	0.9617	0.9715	0.9797	

For the 3 ADIAC subsets, we got approximatly the same results as in the original study, meaning that a high-level feature exctractor like Xception is able to perform as well as case-specific handcrafted features. For the Aqualitas dataset, our evaluation technique of splitting **before** balancing makes our score lower but less biased in our opinion. Finally, the score we got on the Atlas dataset with a significantly higher number of taxa shows that Xception is able to distinguish many taxa with a good confidence.

REFERENCES

[1] DREAL. Atlas des diatomées. http://www.auvergne-rhone-alpes.developpement-durable.gouv.fr/atlas-des-diatomees-a3480.html, 2014.

[2] DREAL. Atlas des diatomées des cours d'eau du territoire bourguignon.http://www.bourgogne-franche-comte.developpement-durable.gouv.fr/atlas-des-diatomees-des-cours-d-eau-du-territoire-a7004.html, 2017.

[3] DRIEE. Atlas des diatomées. http://www.driee.ile-de-france.developpement-durable.gouv.fr/atlas-des-diatomees-a2070.html, 2014.

[4] Pedraza, A., Bueno, G., Deniz, O., Cristobal, G., Blanco, S., and Borrego-Ramos, M. Automated diatom classification (part b): A deep learning approach. Applied Sciences 7 (05 2017), 460.

[5] ADIAC. Public data adiac project. https://rbg-web2.rbge.org.uk/ADIAC/pubdat/downloads/public_images.htm, 2002.

[6] Dimitrovski, I., Kocev, D., Loskovska, S., and Džeroski, S. Hierarchical classification of diatom images using ensembles of predictive clustering trees. Ecological Informatics 7, 1 (2012), 19 – 29.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my Georgia Tech advisor M. Cedric Pradalier and M. Martin Laviale who have been of great support during all the thesis, providing valuable advice and motivation. Besides, thanks to my partner Souhila Founas for her positive energy and great work on this project. Equally, the completion of this project could not have been accomplished without the support of the UMI 2958 CNRS-GT and UMR 7360 CNRS-UL laboratories, providing funds and hardware support. I would also like to thank my INSA Lyon tutor Ms Christine Solnon which have been providing precious feedbacks and help during all my internship. Finally, many thanks to my schools INSA Lyon and Georgia Tech to give me such a great opportunity but also for giving me the knowledge and tools to make it happen.