New framework for analysis of aquatic ecosystems

Ankita Ravi Vaswani and Klas Ove Möller[1]

[1] Biological Carbon Pump group, Institute of Carbon Cycles, Helmholtz-Zentrum Hereon, Geesthacht

Funding period: 6 months (01.08.2022 – 31.01.2023)

Abstract

State-of-the-art underwater imaging systems provide an exciting opportunity to observe billions of individual organisms in their natural habitats at unprecedented spatiotemporal resolution. To unlock the full potential of these advances, we require new analysis pipelines that go beyond classifying organisms by taxonomic groups, and quantify functional traits and biological phenomena from images. Critically, these tools must be made accessible to domain specialists without programming expertise and deployable at scale on modern supercomputing systems. We develop such an image analysis pipeline, manually annotate functional groups, traits and biological processes in images, and train convolutional neural networks (CNNs) to automate and scale analysis of massive zooplankton image datasets. Our pipeline, implemented on a high-performance computing (HPC) system and combining multiple existing open-source frameworks and libraries, provides an intuitive web interface for browsing, searching and annotating images, and allows multiple simultaneous users to work on a single copy of the data online. Images and annotations are then used for both supervised and unsupervised training of convolutional neural networks (CNNs), with the results made available in the web interface. We demonstrate this approach by classifying ~700,000 images to identify functional groups (copepods, diatom chains, *Noctiluca scintillans*, marine snow, etc). Organisms are further annotated for relevant functional traits. Using these trait annotations, future work will further train CNNs for object detection and feature extraction, thereby iteratively fine-tuning CNNs to perform increasingly complex trait extraction from images. We foresee that these tools will enable new avenues of investigation in aquatic research, ecosystem modelling and global biogeochemical flux estimations, revealing previously inaccessible relationships between species biodiversity, zooplankton traits and seasonal variations in environmental conditions.

Introduction

Zooplankton are essential for aquatic food webs and make important, yet incompletely understood, contributions to biogeochemical cycles[1,2]. Understanding how changing environmental conditions affect distribution, abundance and physiology of zooplankton allows us to decode the effects of climate change on biodiversity, ecosystem dynamics and global carbon cycles[3–5]. High-throughput, underwater *in situ* imaging techniques are now frequently deployed to generate billions of high-quality observations of organisms in their natural habitats. This opens up aquatic ecosystems to more detailed study than previously possible. Furthermore, *in situ* images provide information about individual organisms' survival, growth, reproduction and resource acquisition from the visual signatures of the underlying traits[4,7] such as feeding behavior, lipid reserves[8], egg clutch sizes[9], appendage extension[7], and body posture[7]. Hence, underwater *in situ* observation of plankton can be used to extend traditional, species-centric, classification approaches to include information on functional characteristics or ‘traits’ of organisms[5]. Recently, zooplankton traits analyzed during spring ice-melts in the Arctic Ocean have revealed complex ecosystem responses to environmental changes[7].

Analysis of high-throughput *in situ* plankton images by experts is time and effort intensive and is a major challenge for a large-scale analysis of ecosystems[5]. Machine learning techniques are a promising tool for the automated information extraction from hundreds of millions of images, but require large numbers of labeled images for training CNNs[8]. Semi-supervised, machine-learning paradigms allow a performance boost in label-scarce contexts by effectively utilizing the large number of unlabeled images in an unsupervised pre-training step, followed by a supervised, fine-tuning with labeled images[9,10,11]. Active-learning paradigms can further aid effective labeling in highly unbalanced datasets. In this approach, a small, randomly selected subset of labeled data is used to train a CNN. The CNN model is then used to predict classes in the whole dataset, and a confidence or certainty metric reported by the CNN for every prediction is then used to select the least confident predictions (i.e. the most informative images) for further labeling[8].

However, to truly realize the potential of these sophisticated imaging and machine learning techniques, we need a framework that allows domain experts with little or no programming expertise to (1) efficiently store, browse, filter and interact with a large number of images, (2) manually annotate highly unbalanced datasets to generate training data, (3) train CNNs using semi-supervised or supervised machine learning paradigms, (4) evaluate CNN performance to pick neural network architectures and machine-learning paradigms to develop a CNN most suited for a specific task, and (5) implement active-learning paradigms using confidence estimates of CNN predictions to select images for labeling from highly unbalanced data to iteratively streamline analysis.

State of the art

Our pipeline, implemented in a high-performance computing (HPC) system, incorporates an open-source, labeling platform Label-Studio with a PostgreSQL backend that allows multiple concurrent users to browse, filter and label millions of images on a single copy of the data online. Label-studio can be implemented with an integrated ML-backend or labels and annotations can be exported for CNN-training. Using Label-Studio implemented on our HPC, we manually annotated ~ 15,000 images for the purpose of classification into relevant taxon units.

We used a custom-built CNN classifier (Schanz et al. in press), henceforth referred to as ‘Plankton-classifier’ with a ResNet50 feature extractor for label-free pre-training on ~700000 images, followed by supervised fine-tuning with manually annotated labels (Plankton-classifier repository will be made public upon publication of Schanz et al. in press). We also provide jupyter notebook-based tools for formatting annotations for CNN-training, evaluating CNN performance, incorporating CNN predictions back into Label-Studio projects, filtering CNN predictions based on certainty/confidence metrics.

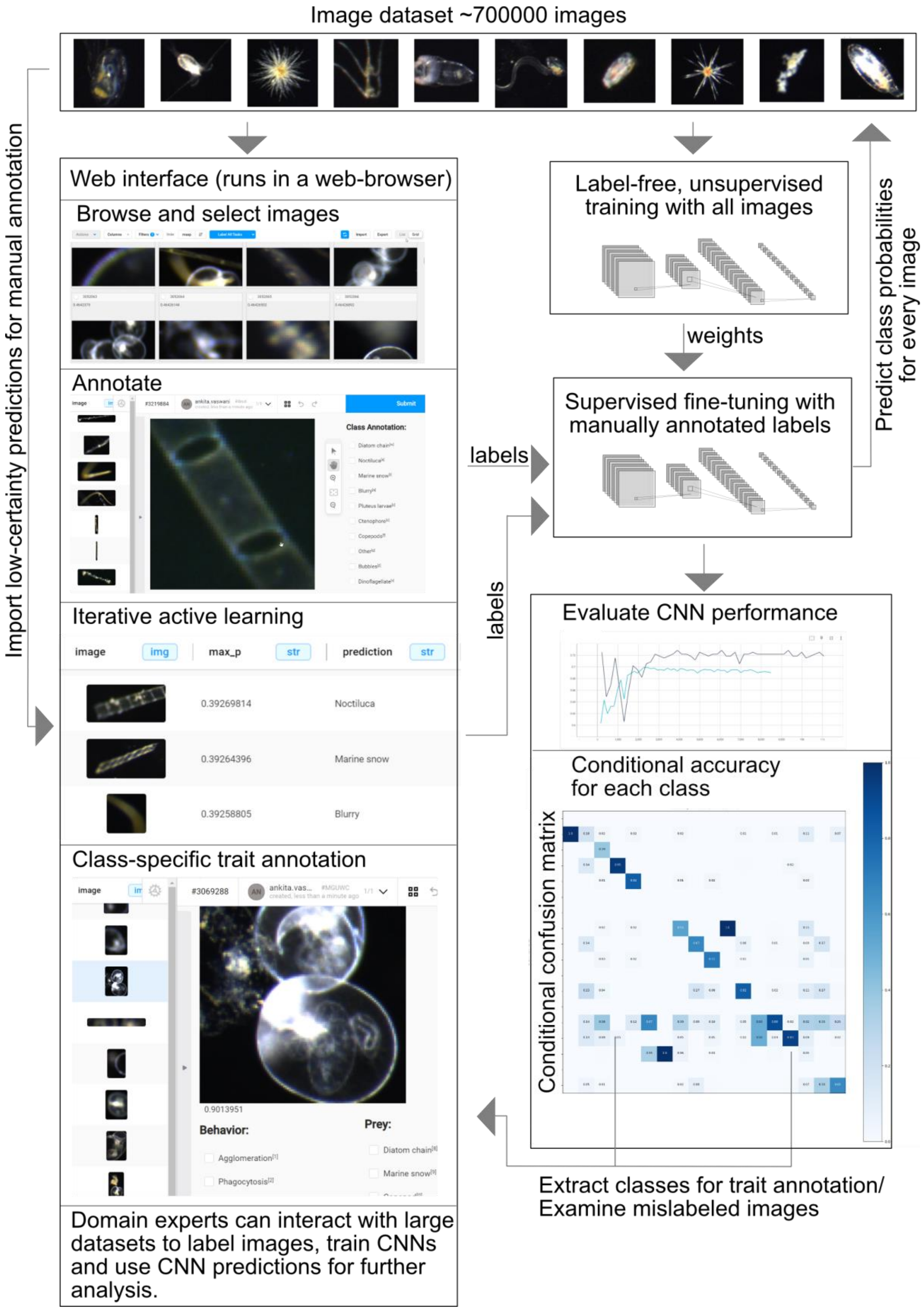
Image dataset ~700000 images
Import low-certainty predictions for manual annotation
Web interface (runs in a web-browser)
Browse and select images
Annotate
Class Annotation:
Submit
Diatom chain
Noctiluca
Marine snow
Blurry
Pluteus larvae
Ctenophore
Copepods
Other
Bubbles
Dinoflagellate
Iterative active learning
Class-specific trait annotation
#3069288
0.9013951
Behavior:
Agglomeration
Phagocytosis
Prey:
Diatom chain
Marine snow
Domain experts can interact with large datasets to label images, train CNNs and use CNN predictions for further analysis.
Label-free, unsupervised training with all images
weights
Supervised fine-tuning with manually annotated labels
Predict class probabilities for every image
labels
labels
Evaluate CNN performance
Conditional accuracy for each class
Conditional confusion matrix
Extract classes for trait annotation/ Examine mislabeled images

image	max_p	prediction
	0.39269814	Noctiluca
	0.39264396	Marine snow
	0.39258805	Blurry

Figure 1: A consolidated data analysis pipeline D1 built to classify *in situ* plankton images, extract morphological features and guide users through semi-automated trait segmentation.

Outcome

1. **Image Analysis Pipeline (D1):**

We provide detailed documentation of our image analysis pipeline (deliverable **D1**, **Figure 1**). Such a consolidated pipeline does not currently exist for trait extraction and will be an invaluable tool for marine researchers. This includes:

1. Implement a labeling interface: While any web-based tool would fit our pipeline, we used an existing, open-source platform (Label-Studio). We provide instructions to implement an instance of Label-Studio with a PostgreSQL backend on an HPC. We include instructions and code to generate file lists for local file-serving, and import into Label-Studio. Detailed templates (HTML tags, CSS styling) interface for plankton classification, CNN evaluation, trait and biological process annotation.
2. Generate training and validation data: Manual classification of randomly-selected images in the implemented interface. We provide documentation and jupyter notebooks to support information exchange between Label-Studio and a CNN (in this case, our custom-built plankton classifier).
3. Train a basic CNN for classification: We used a custom-built CNN (referred to as 'Plankton-classifier) and implemented a semi-supervised machine learning paradigm[10], code for the Plankton-classifier will be available upon publication (Schanz et al. in review).
4. Use CNN performance metrics to evaluate and choose between machine learning paradigms and hyper parameters. We provide out experiment results and documentation to enable researchers to choose the machine learning approach.
5. Run the CNN model on the entire dataset to predict classes or functional groups with a high conditional accuracy.
6. Further manual annotation of visual signatures of functional traits for class-specific trait annotation.

2. **Annotated class and trait datasets (D2)**

We applied D1 to images acquired during research expeditions in the North Sea to produce an annotated class-labels dataset (~6000 images, Kordübel et al. in prep). We trained our Plankton-classifier on these labels, evaluated CNN accuracy on withheld labels and used the CNN to infer predictions on the entire dataset (all ~700000 images, **Figure 1**). To each image, the Plankton-classifier assigns probabilities for each class, and the class assigned the maximum probability (max_p) is the predicted label. We then selected all images with max_p < 0.4 (~7000 images) for manual labeling. We provide ~14000 class annotations (images will be made available upon publication) and their max_p values as the deliverable **D2**.

The Plankton classifier predictions were used to extract classes with high conditional accuracy such as *Noctiluca*, diatom chains, marine snow, etc. (**Figure 1**). We are currently generating relevant trait annotations for these classes and will add the trait annotations to the deliverable D2 upon completion.

3. Plankton-classifier CNN (D3)

All the labels generated (D2) were used to train the plankton classifier (CNN, **D3**) in collaboration with Hereon's Model-Driven Machine Learning group (MDML). The code repository **D3** will be made public upon publication (Schanz et al. in review).

4. Code Repository (D4) and Jupyter notebook tutorial (D5)

Documentation and code for our data analysis pipeline (D1) are provided in the current version of our code repository (D4). Upon publication of the Plankton-classifier (Schanz et al, in press), we will provide a Jupyter notebook tutorial (D5) of our data analysis pipeline to guide users through annotation for classification, trait segmentation, CNN training and data visualization.

Outlook and Summary

In collaboration with the Helmholtz AI Cooperation Unit (Helmholtz AI), we are currently working on incorporating a conversion to binary formats to scale our data analysis pipeline to deal with larger datasets (~10^8 images). Additionally, we are generating trait annotations that capture information about relevant characteristics or morphological and behavioral properties in images. These annotations will be used train a CNN for automated object detection and feature extraction, in collaboration with the MDML group at Hereon. To improve accessibility by non-programming, domain experts, we plan to develop a GUI for our data pipeline.

Automatic taxonomic classification and trait extraction (D1/D3) will be valuable for marine biologists, ecologists and image analysts. We hope that the tools developed here will enable domain experts in aquatic research, ecosystem modelling and global biogeochemical flux estimations, to analyze previously inaccessible relationships between biodiversity, zooplankton biology, seasonal variations in environmental conditions and impact by climate change.

References

1. Frederiksen, M., Edwards, M., Richardson, A. J., Halliday, N. C. & Wanless, S. From plankton to top predators: bottom-up control of a marine food web across four trophic levels. *J Anim Ecology* **75**, 1259–1268 (2006).
2. Boyd, P. W. Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* **9** (2016).
3. Violle, C. *et al.* Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).
4. Ohman, M. D. A sea of tentacles: optically discernible traits resolved from planktonic organisms in situ. *ICES Journal of Marine Science* **76**, 1959–1972 (2019).
5. Orenstein, E. *et al.* Machine learning techniques to characterize functional traits of plankton from image data. *In press* (2022).
6. Hatton, Heneghan, Bar-On & Galbraith. The global ocean size-spectrum from bacteria to whales. *Sci. Adv.* **7**, (2021).
7. Vilgrain, L. *et al.* Trait-based approach using in situ copepod images reveals contrasting ecological patterns across an Arctic ice melt zone. *Limnol Oceanogr.* **66**, 1155–1167 (2021).
8. Bochinski, E. *et al.* Deep Active Learning for In Situ Plankton Classification. *Pattern Recognition and Information Forensics.* **11188,** 5–15 (2019).

9. Lumini, A. and Nanni, L. Deep learning and transfer learning features for plankton classification’, Ecological Informatics, 51, 33–43 (2019).

10. Chen, T. *et al*. A Simple Framework for Contrastive Learning of Visual Representations. arXiv: 2002.05709.

11. Schanz, T., Möller, K. O., Ruehl, S. & Greenberg, D. Robust Detection of Marine Life with Label-free Image Feature Learning and Probability Calibration. *Helmholtz AI Conference, Dresden* (2022).