Ocean Sci. Discuss.,
https://doi.org/10.5194/os-2020-62-RC1, 2020

Ocean Science
Discussions
Open Access

Interactive comment on “Towards operational phytoplankton recognition with automated high-throughput imaging and compact convolutional neural networks” *by* Tuomas Eerola et al.

Anonymous Referee #1

Received and published: 4 August 2020

GENERAL COMMENTS

The manuscript addresses automated classification of brackish-water phytoplankton images collected in the Baltic Sea. Images used in the experimental part were collected using the IFCB (Imaging FlowCytobot) system.

There is an extensive discussion regarding current state and challenges around plankton image classification, including issues such as the need of Computer Vision and Machine learning based approaches to deal with the large volume of data, limitations

of traditional machine learning based approaches that require hand-designed feature engineering and therefore do not scale well from one case to another, recent advances towards deep learning based approaches as they are able to automatically extract features from raw images, the need of large amounts of training data, imbalanced class sizes and the need to deal with open classes or unconstrained class (e.g., detritus), and so on. There is also a clear concern with respect to practical implications of using these classifiers.

Effectively, a reduced scale size CNN network is proposed and trained from scratch, varying some training conditions, for datasets in the above Baltic Sea context. Classification accuracy up to 0.85 is achieved and observed confusion cases are extensively discussed. The proposed network and performed experiments are apparently meant to show that a small scale CNN can be trained from scratch on limited and imbalanced training data condition, using data augmentation strategies. This agrees with what is stated in the Abstract: "Our results show that it is possible to obtain good classification accuracy with relatively shallow architectures and a small amount of training data when using effective data augmentation methods even with a very unbalanced dataset."

These issues, training small networks from scratch, using data augmentation, dealing with imbalanced classes, have been already at some extent addressed in previous publications on plankton image classification. Motivating issues as well as discussions regarding challenges of making effective practical use of these classification methods are also discussed in previously published papers. Since no comparison is provided, it is not clear what are the contributions of the presented method and experimental results.

In summary, the objectives and the contributions of the study are not clear.

SPECIFIC COMMENTS

It is stated that "Our approach ... is to address some fundamental challenges in phytoplankton identification". What are precisely these fundamental challenges?

A convolutional neural network architecture is proposed, trained, and results are presented. Some results refer to input of size 128x128 while others to input of size 256x256. However there is no comparison between them nor with any other existing models or datasets. This makes very difficult to evaluate the relevance of the study. In particular, since there are many lightweight models being proposed for image classification tasks in general, one wonders why none was considered or at least used as a reference.

There are several reports, not only with respect to plankton images, of how transfer learning (models pre-trained on images of a distinct domain and fine-tuned with target domain images) often generates classifiers that achieve good classification rates. It seems important to compare the results achieved with the proposed network and results that could be achieved by fine-tuning pre-trained models. Note that models pre-trained on ImageNet data are often used, but any model could be also pre-trained on any large datasets (e.g., on large plankton datasets).

Dataset description in the "Materials and methods" section is confusing. As stated, and as listed in Table 1, it consists of 53 classes. Then it is said that they are further subdivided in subclasses that results in a total of 61 classes. Where or how this subdivided case is explored ?

Experimental setup: The first main issue is what the experiments are trying to convey (with respect to current state-of-the-art). Then, there are some details related to experimental setting that needs clarification: (1) if evaluation is performed based on cross-validation, why the dataset was separated into training-testing sets (25% for testing) ? (2) When performing cross-validation, did you consider stratified folds? (3) It is stated that "The parameters are based on small-scale empiric tests where it was observed that the CNN can be trained successfully with these parameters." What kind of empirical tests are they? Did you take care so as to not introduce any bias (parameters chosen on privileged information) ? (4) When comparing CNN with RF, it seems that different cross-validation fold numbers have been considered. To avoid performance

differences that may be due to fold differences, the same folds should be considered for training/testing of both algorithms. (5) In general, establishing a baseline case helps comparison; for instance, as results are presented, it is not clear why some results refer to CNN128 and others to CNN256. How they compare each other?

TECHNICAL / SMALL DETAILS

"collaboration between experts and exchange with other disciplines, like modelers". I did not understand what is the meaning of "modeler" here.

"The number of images in a subset assigned to the testing set is equal to 25% of the threshold value of the subset. The remaining images, up to one thousand images, are then assigned to the training set." Do you mean "up to one thousand images PER CLASS" ?

Table 2: does the number of test images per class refer to the smallest class? If so, separation of 25% for testing was class-wise ? (but still, it is not clear where this division is considered)

Data augmentation: with 90-degree rotation, we have 0-degree (original), 90-degree, 180-degree, and 270-degree. So it would be a 3x augmentation and not 4x, right ?

"Validating any results related to class specific classification accuracies becomes computationally expensive since there are few images in a moderate number of classes". This sentence is confusing for me. If there are many images and large number of classes, I would understand that evaluation may become computationally expensive, but the other way is not clear.

Table 3 and Table 4: description of the architecture is not following the standard. For instance, when specifying the dimensions of a tensor, the standard is h x w x d (height x width x depth). Also, since you are using non-usual filter sizes (masks 10x10, 6x6, 4x4), their choice should be justified.

Batch size is a hyperparameter that can greatly affect convergence. There are some recommendations to use relatively small batch sizes (32 or 64) or to reduce it along the epochs. It could be interesting to evaluate different values.

Which loss function was used? Did you use any kind of regularization other than dropout ?

The first time I read the part that mentions cross-validation, I have understood 30,30,30, 60 were referring to the number of folds. But at a second reading, it seems to refer to the number of repetitions. Could you precisely describe how you did each cross-validation ?

Since computational cost is a concerning issue, which type of processor has been used and how long was the training time ? What is considered a "short training time" ?

Would it be possible to display results in a confusion matrix ? At least the more interesting cases ? It is difficult to follow the results in Tables, as shown.

Image size is mentioned as a metadata used in the experiments. If properly cropped, image size could reflect organism size in terms of pixels. However, for this information to be useful, size should correspond to the physical size of the organism. Estimating the actual physical size may not be so simple if we do not have precise distance information.

"typical CNNs struggle in open-class problems where the method is applied to novel data with classes not present in the training data". Is this not true for most of the machine learning algorithms ?

Interactive comment on Ocean Sci. Discuss., https://doi.org/10.5194/os-2020-62, 2020.