Qeios

Peer Review

Review of: "WOLO: Wilson Only Looks Once – Estimating Ant Body Mass From Reference-Free Images Using Deep Convolutional Neural Networks"

Tom Kirstein[1]

1. Institute of Stochastics, Universität Ulm, Germany

Overall, I think the manuscript is of very high quality. The described approaches are rigorous and methodologically sound. In my opinion, training the networks using synthetic image data of ants is very interesting; I've used similar approaches for synthetic microstructures but was not aware that such approaches can be extended to something as specific as an individual ant species. My expertise is in evaluating the machine learning aspect of this paper; however, I do not have expertise in biology and can thus not evaluate the impact of this work.

Despite the general quality of the manuscript, there are still areas that could be improved or clarified.

>As a consequence, global cues are usually vital for robust size estimation, but they cannot always be provided and, at the very least, limit application versatility

Can you provide specific examples or scenarios where reference-free size estimation would be more advantageous than using a reference model for ants?

>Regression was conducted on 128 128 3 image samples (resolution in x, resolution in y, colour channels)

This may be a display error, but I would prefer "$128\times128\times3$ if possible.

>Regression is arguably the most natural implementation of the mass-estimation problem, but suffered from prediction bias

I would agree with the authors that regression is the most natural implementation for size estimation, but I do not understand why it would introduce a bias when trained on the MSE. Did the authors

doi.org/10.32388/L2LPZR

investigate where this came from? (The authors later discuss this, but the observed results are not a bias in the classical sense)

>A reasonable alternative is to minimise the relative error, realised by training networks on log10-transformed body masses (Fig. 3D).

Did the authors train the network using such relative errors (such as MAPE and SMAPE)? These would, in my opinion, be more natural choices than the MSE. The authors mention later on (this could, in my opinion, be moved up) why the MAPE wasn't used, but the question remains why the SMAPE wasn't used for training.

>This bias is likely the result of setting the absolute error as the loss function

I thought the mean squared error was used?

>A key difference between classification and regression is that all classification errors are equal

This sentence is, in my opinion, not correct. In many classification approaches, such as the one laid out by the authors, classification errors can be weighted differently. The above sentence should therefore be adjusted.

In the results section, I felt that it was difficult to follow what exact data was used during the training of the network. In particular, I expected the synthetic data to be used, but this was only done in 3.5 onwards? I understand that additional information is provided in the supplementary, but a more explicit statement of how exactly the training is conducted (i.e., what data is used for training) in the manuscript could help the reader. In the results section, I also asked myself how much of the misclassification error originates from those examples for which the ground truth size is close to the boundaries of two size classes.

>Mass predictions were almost four times more accurate for the largest compared to the smallest workers, a result most clearly visible when the regressor output is translated into a 20-class classifier equivalent, as shown in (B).

To me, it looks as if the smallest class has an accuracy of about 0.6; the large drop-off occurs at the second smallest size class.

>Accuracy bias was substantially weaker for regressor networks trained on log10-transformed body masses

As mentioned above, I would not use bias here. I suggest replacing it with ‘accuracy variation’ or something like that to avoid implying a systematic error across all classes. Consistency in terminology (e.g., ‘prediction bias’ vs. ‘accuracy bias’) would also help.

Writing style is mostly great; some small improvements could be made:

>The prediction stability and precision of predictions for continuous variables is best assessed via the coefficient of variation (CoV)

Double prediction at the start could imo be simplified to “The stability and precision of predictions”.

>However, no CoV can be defined for classifiers, so rendering comparison between regressors and classifiers impossible.

Remove the so.

Declarations

Potential competing interests: No potential competing interests to declare.