Qeios

Peer Review

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

Paul Tresson[1]

1. Institut de Recherche pour le Développement, Marseille, France

I've appreciated reading this paper on reference-free size estimation in ants. The integration of multiple tools and techniques to overcome the technical challenges displayed is very interesting, and the underlying science seems to be sound.

I do have some comments, however. Indeed, some important points need further clarification. Most importantly, while the different datasets used are described thoughtfully, the ultimate train-test procedure is not as clearly described as it should be. This point is paramount to be able to evaluate the quality of any work in ML and even more so in DL.

As such, it is therefore hard to evaluate the reported performances. The accuracy of 11% is reported, but if my understanding is correct, this is only achieved on a dataset very similar to the training dataset and not directly usable as such? This must be clarified.

One thing I've felt missing, also for a work working between living things and DL, is the distribution of classes in the training and testing datasets. Again, if my understanding is correct, weight classes are balanced in the datasets presented here. Is it so in the wild or in the lab? If not, this must be taken into account. Class imbalances are, in my opinion, one of the most common challenges in any DL work applied to living things - where things are not uniformly distributed.

Lastly, the authors seem to be missing several classical deep learning techniques to improve performances. Why use a VGG when it is far from SOTA for many tasks? Why not use pre-trained weights? Why not more data augmentation? All these techniques are classics that have proven to be useful in the last decade; why pass up the potential to greatly improve model performances for a

 doi.org/10.32388/46FWBL

relatively low effort? I use PyTorch and not Keras, but I believe that using pre-trained weights is one argument away? As is using a ViT, a ResNet, or others...

More generally, an open and possibly naive question: on natural images, a clue for the size of things is often the focus or lens effects (images of insects typically are sharp and focused on the subject with a blurry background). Could lens distortions or such also be a learnable clue for the size of a subject? Then, is this factor accounted for with generated images?

Also, a remark that is not directly related, but it is very pleasing to see the integration of research tools within Blender and using Unreal Engine.

Detailed comments

1. Introduction

- "However, completely reference-free size estimation is rare[6]."
Could you provide more examples and use cases where this might be needed? Maybe it's rare because you'll often have an easier and more reliable solution to measure something (e.g., standardization and such)?
It is easy to imagine that you are not the only one needing to work on reference-free images, but could you provide more general motivation on why and when this is a problem to solve?

- "To give but a few examples, size frequency distributions of foraging parties appear to be adapted to the spec## Abstractific requirements of the available food sources[14][33], and are affected by food source structural and mechanical properties[14][17][22][34][35]."
This is a bit unclear; maybe you could provide a concrete example to illustrate your point? But this is more to clarify than anything; if this is possible in a short sentence, I feel like it would be welcome to give context to the reader; otherwise, this is not needed.

Maybe flesh out more the content of the article in the last paragraph of the introduction?

2. Methods

2.1. Data collection and curation

2.1.1. Training and Benchmark Datasets

If I understand correctly, there are only 20 individuals recorded in the MultiCamAnts dataset? Does this not lead to strong correlation between images of this dataset?

"Leaf-cutter ants to represent 20 body mass classes that cover the worker size range of 1-50 mg; class centre-to-centre distances were spaced equally within this range in log10-space to achieve a more fine-grained class resolution among more common smaller worker sizes"

This sentence is confusing; maybe split it into two?

How was the split done between train and test samples? This is not directly cleared up in the methods section.

You mention "and split into 2.5 million training and 0.5 million validation samples (80/20)."

I deduce that Test-A and Test-B datasets are independent testing datasets never seen in training. In this case, these methods seem to be reliable.

Otherwise, if you only perform a random split, with your setup, I fear that the test set would be highly correlated with the train set and that this would lead to overfitting!

In any case, this needs to be made absolutely clear on first reading (just by adding a sentence at the end of the first paragraph).

With the orders of magnitude between different workers and the "relatively" low number of ants actually weighted (I do understand the effort that went into the setup of these datasets), maybe it would be good to provide some measurement accuracy confidence in your balance or something? How sure are you that the ant on the top right corner of Fig.1 weighs 35.8 mg?

Also, what is the expected weight distribution in the wild? I strongly suspect it is not uniform, and your dataset seems to be. Your training dataset must reflect the conditions of usage afterwards! If in the wild (/ant farm in the lab) there are 100 class 0 workers for 1 class 4, I fear your model would be unusable. Do provide details on this topic...

2.1.2. Augmentation with synthetic data

The augmentation with synthetic data is a very interesting approach. I'm always curious to see the effects on the robustness of a model.

Did you implement some noise variation in the masses, or was it always the precise steps in Fig.2?

2.2. Inference approaches

This section describes training more than only inferences; maybe update the name?

2.2.1. Regression

Did you test other (more recent and closer to SOTA...) architectures? A lot of more recent networks than VGG perform better, even on very niche tasks. I feel like this choice needs to be motivated and argued.

"Regression was conducted on 128 128 3 image samples" -> RGB images were cropped to 128x128.

Were there additional data augmentations and transforms (I expect normalization, but maybe also flips, rotations, color shuffles, etc. could be possible?)

You don't seem to use pre-trained weights; why not?

Did you consider other losses (MSLE, Huber, quantile loss...)? The accurate loss does also depend on the distribution of your data!

Can you provide a train/test loss curve in SM?

2.2.2. Classification

"20 classes roughly match the discretisation used by Wilson" What does “roughly” mean?

The implementation of label smoothing seems to be a relevant idea!

2.3. Evaluation

2.3.1. Prediction error and accuracy

It is a good thing to note the inherent error baked into the classification approach.

Did you compute other metrics for regression, such as R^2 between pred and obs?

2.3. Evaluation

Independent evaluation and comparison with human performances is a very good idea.

Why include colleagues without experience? How many were experienced or not? (Table S3 does not seem to match a description of the participants?)

14 colleagues represent a very small sample; I do not believe that the metrics could really be statistically significant. I imagine that specialists in leaf-cutter ants body size estimation are hard to come by... Maybe this whole part only belongs in the discussion as food for thought and perspective? Even if not statistically significant, this is a nice experiment that deserves a mention and brings completeness, in my opinion.

3. Results and discussion

Why not have separate results and discussion ? The results are harder to find this way I find.

3.1. Regressors achieve intermediate prediction errors, and can be strongly biased

Fig. 4 "The most natural implementation of the mass estimation problem is a regressor network.". This sentence is unnecessary.

3.2. Classifiers achieve higher categorical accuracy, but suffer in prediction spread

idem with Fig. 5. "Classification presents an alternative method to size estimation" belongs in the discussion

3.4. Performance on out-of-distribution data

Are networks with these error rates really usable in practice ?
I feel like the work beginned here is of good quality and relevant but in the end this is the relevant metric... How well does these networks perform on new unseen data in the wild/lab ?
There is still a lot of room for improvement (starting with using pre-trained modern networks). I am confident that much better results are easily achievable.

3.5. Synthetic data increase prediction robustness

For me this is maybe the most interesting result in this study ! I would have loved more details. For instance, at which point does the addition of CGI data becomes to be detrimental ? How far can you go from acquired data and still improve robustness ?

3.6. The best networks outperform humans

As *mentioned* before, even if these results cannot be interpreted as significant, this is still worth noting.

I believe the following sentence to be indeed accurate : "This pilot study thus cannot provide a reasonable indication of the upper limit of human performance; but we believe that it supports the weaker conclusion that the task itself is hard"
the following "as well as the assertion that the trained networks learned a considerable amount from the training data provided." seem to be an overstatement however, I would cut that part. Or make a more neutral version "confirms that this task is indeed learnable.." or something like that ?

It is very interesting to compare human vs. computer errors ! And conforting as they are similar.
Could you provide a correlation between human and computer estimation ? As to have a quantified metric rather than only a couple of examples.

It would also be interesting to study the outliers : the images correctly predicted by humans and not computer and vice-versa.

4. *conclusion*

Is the 11.94% accuracy reported the real accuracy of the network ? I would consider the performances on Test-B the real accuracy.. Then is this really usable as is ?

There is a lack of perspectives in the discussion and conclusion. What concrete things could be done to improve performances in future research ? What could be other applications ? How does it relate to the evolution of the state of the art in computer vision ? What possibilities does it unlock in entomology and ecology as a whole ?

Declarations

Potential competing interests: No potential competing interests to declare.