Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-196-RC2, 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "GRAINet: Mapping grain size distributions in river beds from UAV images with convolutional neural networks" *by* Nico Lang et al.

Patrice Carbonneau (Referee)

patrice.carbonneau@durham.ac.uk

Received and published: 19 July 2020

Review of GRAINet

This paper develops a deep learning (DL) workflow to map grain size. The method is innovative because it fits the network directly to a distribution. The results seem extremely impressive and promising. But I have several major concerns that may need to be addressed before the work is ready for full publication.

Major issues

Drone Flights and Image resolution

The flight design for the method is rather inconvenient. Data is captured by imagery at an altitude of 10m with full 80% overlap for SfM. What the authors don't explicitly mention is that this leads to very long flight times to cover even small bars. This is not the norm. Many UAV users (and there are many of us) will read this paper wanting to apply the method to our fieldwork. But most drone acquisitions are done at 50+ meters. So in essence, the method as described here is a DL equivalent to Robotic Photosieving (cited in the work). The method would have much more impact if it could operate on the existing flight patterns (50+ meters flight altitudes) and show that you don't need the near-ground flights to get grain size distribution. The authors did try to look at the effect of resolution. But in my experience with image texture work, I have repeatedly found that downsampling an image digitally is not the same thing as acquiring the scene from a higher altitude. Such operations are approximations at best. As presented the method is an interesting alternative to Robotic Photosieving, but it doesn't have the transformative effect we have come to expect from Deep Learning methods. In section 5.45, the authors claim that it is novel to have dense grain size maps at scales of 1.25m by 0.5m. This is incorrect. Carbonneau et al (2004), which is actually cited here, produced grain size maps using image texture calculated on a 33x33 pixel window for 3cm imagery, the scale was therefore 1mx1m which is arguably very similar to the scale here. In 2005, the same group of Carbonneau et al showed that this could derive maps of D50 for en entire river of 80Km and presented grain size profles based on several million image-based measurements of D50. The authors should then consult Woodget, Fyffe and Carbonneau (2018) in ESPL where we present a further image-bsaed method for particle sizing. For the drone th authors are using, operations at 50m AGL will deliver imagery at $\sim$2cm. This means that a texture based approach can deliver D50 estimates for sub-m2 patches. The authors need to choose their wording more carefully. The innovation here is that for each patch, GRAINet gives a distribution. But in terms of mean diameter, this has been possible for more than 15 years. Even in terms of the distribution, if a user applies the Robotic Photosieving method and spends as much time in near ground acquisition, then distributions can be derived from a large number of images. We have found that the new PebbleCounts algorithm (Puriton and Bookhagen, this journal) is much faster and more accurate than Basegrain. This will get distributions for each image. This means that the progress of GRAINet is in processing time and perhaps better precision. But the issues with generalisation mean that users will need to train the network for new rivers. Ultimately, the real innovation potential here would be to show that this method works on a more 'standard' UAV orthomosaics. This would really position the paper as a major innovation on existing methods.

Network Architecture Description

The authors provide a qualitative description of their architecture, and a very detailed description of their loss functions, But the actual architectures they use remain very opaque. See below my point on code availability. At the very least the work needs a figure to detail the network architecture. This is very common in the DL literature. Why this network with residua blocks? Why not just a VGG16 or 19 model with a dense top terminating in a regression layer? Or any other model? Which libraries are used? Even this detail is lacking. It would also be helpful if the authors could show some of the activation maps in order to confirm that the network is 'seeing' reasonable elements that can conceivably lead to a regression to grain size distributions.

Training and Overfitting

In a wider perspective, this is a medium network with in excess of 1 million parameters. Since the classic Goodfellow textbook recommends 5000 samples per class in a classification problem, the sample size presented here does seems small even if I fully realise that much work went into it's production. But it then follows that the authors must establish that their network is well trained. As it stands, there is even a footnote saying it is overfit. Readers less familiar with DL might not realise it, but this is a very serious problem. The authors need to provide the reader with some crucial re-assurances and at least show a tuning plot as seen in figure 1 (from supplementary

information in Carbonneau et al 2020 in ESPL early view).

Validation

The authors maximise their small dataset and use k-fold cross-validation. This is fine to start. But what they call 'geographical cross-validation' is nothing else but bar-scale boot-strapping (sometimes called jack-knifing). It is not appropriate or reliable in this context. The main issue is that even of the authors hold out a whole bar, there remains samples from the same river. Given that river properties vary relatively slowly with geology and sometimes tributary inputs, the boot-strapped training samples will still have data that is very similar to the label data. So this is not a good test of generalisation. A much better approach to this would be to hold out an entire river. If this does not give good results, then the authors must sell this approach as river specific, and consider how much time a new user would need to spend in collecting and producing data to train a new GRAINet. As it stands, I think the results show that the network is not well trained, has too many parameters or too little training data.

Comparison to past work

I have no doubt that this deep learning approach will do better approaches than old texture-based approach, but the authors must be more direct in their discussion when the describe their outcomes. They must show that this is not just a new way of doing existing jobs, it represents real progress. And to do this they must cite errors from other work and clearly demonstrate that their results are better. This should include SedNet which is the rival CNN approach. And given that the title has the term UAV, they must be much more explicit in discussing their method in relation to other UAV (and airborne) particle sizing methods. Ideally, they should try to adapt the method so that it represents real progress with respect to other UAV methods in terms of time spent in acquisition, pre-processing and final data quality.

Code

Perhaps I am missing something but I cannot find the code associated to this work. If this is not a mistake on my part, I view this as a critical limitation of the work. As the authors surely know, the ethics of the deep learning community are strongly oriented towards the open source model. There are countless GitHub sites describing CNN architectures. If Google and Facebook can open-source Tensorflow and Torch, I cannot understand why the authors cannot make their code transparent. In the case of this work, the opaque architecture descriptions make this even more important. Scientists are increasingly criticised for a lack of transparency and something we can and should all do is release the code for such methodological contributions. I leave it to the Editor to decide how important these thoughts should be as this is ultimately a Copernicus policy decision. But I think the time where such methods could be unseen has passed.

Minor issues I have attached an annotated manuscript.

Overall recommendation This is an exciting method that potentially represents a major shift in our ability to measure grain sizes from UAVs. But the work needs major revision before full publication. Key issues:

1. Did the authors not conduct a more conventional UAV survey at higher altitudes when they acquired the data? It would really strengthen the paper if they could show that the method is applicable to existing UAV-derived orthomosaics of rivers sediments now routinely acquired by a large number of researchers. Do higher altitude images even have the required information content to infer distribution? 2. Provide a clearer description of their architecture along with a figure. Indicate libraries they are using. 3. Redo the 'geographic cross validation' to have a proper out-of-sample test set by setting aside entire rivers. 4. Improve and clarify the training procedure with an emphasis on showing that their re-trained models are not overfit. Unfortunately this may require the reprocessing of the data, it's not acceptable to have a footnote and say that your models are overfit, that means they are not ready for publication. 5. Provide a batter comparison with past work and be sure that literature cited in this work has been fully read. 6. Seriously consider a more transparent approach with open-source code.

Patrice Carbonneau July 2020

Please also note the supplement to this comment:
https://www.hydrol-earth-syst-sci-discuss.net/hess-2020-196/hess-2020-196-RC2-supplement.pdf

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-196, 2020.

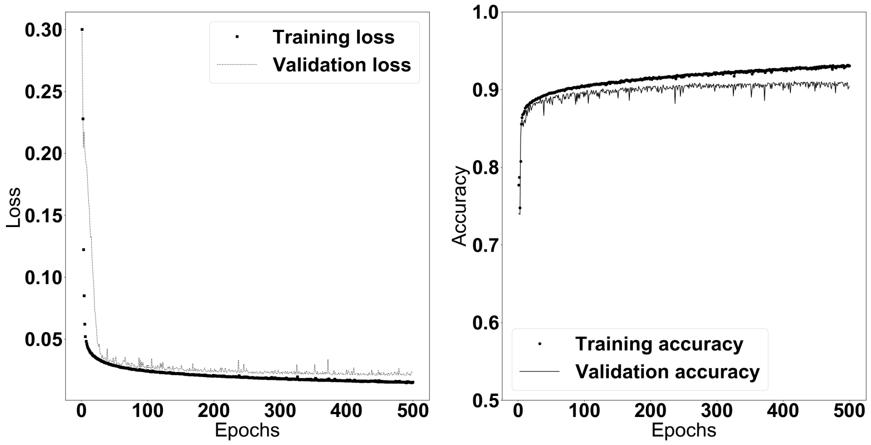
0.30
0.25
0.20
0.15
0.10
0.05
Loss
Training loss
Validation loss
0
100
200
300
400
500
Epochs
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
Training accuracy
Validation accuracy
0
100
200
300
400
500
Epochs

Fig. 1.