The Cryosphere Discuss.,
https://doi.org/10.5194/tc-2020-310-RC2, 2020

The Cryosphere Discussions

Interactive comment on "Image Classification of Marine-Terminating Outlet Glaciers using Deep Learning Methods" *by* Melanie Marochov et al.

Anonymous Referee #2

Received and published: 18 December 2020

General Comments

This paper describes a two-phase deep learning approach for the image classification of Greenlandic marine-terminating outlet glaciers. Optical Sentinel-2 imagery acquired in 2019 over Helheim Glacier was used to train a VGG16 model generating training data for the multilayer perceptron/cCNN in phase two. The results were tested on two Sentinel-2 scenes over Helheim Glacier and Scoresby Sund for summer/autumn 2019. The novelty compared to previous studies is the classification of satellite images into seven different classes. Further results of this study include the performance testing on different tile sizes, transfer learning, and band combinations.

The manuscript is well written and explains the study approach in every detail. There are some concerns regarding the methodical approach and testing of the algorithm

which are explained below. Therefore, I recommend a revision of the manuscript before publication.

As outlined above, some major concerns exist regarding the following points:

- The validation labels include unclassified areas especially at the boundaries between two classes. Why was that done? What does this mean for the accuracy assessment? It seems the accuracy was only calculated over areas where a classified validation label exists. But this approach would miss out on the accuracy over regions with boundaries. Additionally, if no boundaries between classes exist in the training data I would expect that model predictions are inaccurate in those regions. Moreover, your accuracy assessment cannot account for that as validation labels include no data areas. Could you please explain how you handled class boundaries for training and validating the model?

- Testing was performed on only two images acquired temporally close to the training data. Training data was used for 07/08/2019, 01/09/2019, and 28/09/201 and tested for 13/09/2019 and 01/08/2019. This means between testing and training only one to two weeks elapsed. Hence, spectral properties of the images as well as the conditions at the glacier terminus were very similar in the training and test data and might overestimate the accuracies. To show that your approach is transferable in space and time I would recommend testing on a broader amount of images as it has been done by previous studies (Baumhoer et al., 2019; Cheng et al., 2020; Mohajerani et al., 2019; Zhang et al., 2019). I would recommend taking data from a previous/later year e.g. 2018 or 2020 not close to the training data for additional performance testing.

- What was the argument to create a two-phase deep learning model instead of using a fully convolutional network (FCN) architecture for semantic segmentation? It would be great if you could include some comments on that in your manuscript.

- Could you provide some information on the computational cost of the here presented two-phase model compared to semantic segmentation approaches?

Specific & Technical Comments

P2L60: For clarification, it would be great if you could give some more detail on the difference between semantic segmentation by an FCN and the pixel-based segmentation performed here. If I understood you correctly, your first CNN performs image classification, hence assigning one class to the entire image. The second cCNN performs a classification for each pixel. If patch size 1x1 is used, only the spectral properties of one-pixel are used for the classification. In the case of bigger patch sizes, also information on textural features of neighboring pixels can be used for the classification. But semantic segmentation by an FCN would also consider the spatial relationship between pixels of different classes which your approach does not.

P4L117: I think the spatial transferability is not yet proved by only testing on one outsample scene. For applications elsewhere in Greenland and Antarctica more spatially diverse training data would be required. Please mention that or show on a more diverse test image set the spatial transferability of your approach.

P9L266: How did you differentiate between snow on ice and snow on rock?

P11L311: Please describe the term “class raster” more precisely.

P12 Figure2: What does the 1x1 median filter do? Please describe.

P14L348: I would expect that the optimal hyperparameters (epochs, batch sizes, learning rate, etc.) for training are different depending on tile size. Did you experience that?

P15L396: What is the fourth dimension of your 4D tensor? Only three are listed.

P16L420: Please explain the normalization by 16384 in detail. Usually, the min/max values (normalization) or mean/standard deviation values (standardization) of the data set are used for scaling input images.

P16L429: It is not true that your dataset is larger than the previously used ones. The number of tiles might be higher as you use small single class tiles but the number of

images (13) is less than from Zhang et al. 2019 (75), Cheng et al. 2020 (20188), and Baumhoer et al. (38 scenes). You state this on P29L878.

P28L840: Be careful with comparing your F1-score directly with the one of Xie et al. (2020) and Baumhoer et al. (2019). Both studies used a more diverse set of test data. Moreover, Xie et al. calculated the accuracy also over the boundary between two classes and this is the area where errors occur. Additionally, Baumhoer et al. performed their accuracy analysis on a 1 km buffer at the calving margin to account for inaccuracies at the frontal area, where again, the inaccuracies occur. P28L849: I guess for future potential applications (e.g. glacier terminus tracking, snow line extraction, coastline mapping, etc.) especially the edges between classes are of major importance. Is it possible to get clear class boundaries from your classification result?

P29L898: You are right, that optical imagery is easier to pre-process but please also mention that SAR data has many advantages. Especially in polar regions, optical data availability is very limited due to cloud cover and polar night. SAR data overcomes those drawbacks and allows continuous time series with plenty of data.

P30L902: Be again careful with not confusing tensor size with the number of input channels.

P30L912: Maybe re-phrase or delete this sentence. Arguing by the number of bands whether a model mimics human visual performance is confusing.

P30L915: The paragraph comparing your classification approach to the U-Net architecture is slightly misleading. The U-Net allows semantic segmentation of images by delineating features. The U-Net learns shapes and forms but is not limited in variability unless the training dataset is restricted by too little data and missing augmentation. In natural landscape images, the challenge of color is often given by the fact that two classes (e.g. snow on ice and snow on rock) have similar spectral reflectance but a different texture and/or shape. That is why the U-Net is so powerful as it also considers the spatial context besides pixel values. To show that your approach exceeds the U-Net

architecture you would need to prove that it is as suitable for delineation on a larger test set. Therefore, I think it is problematic to conclude that the “compact CNN architecture has exceeded the results from the U-Net architecture”. Your approach concentrates on the pixel-based classification of classes (and was tested for that) whereas the U-Net based approaches concentrated on the correct delineation between classes.

P33L1028: Again, I would be careful with class boundaries unless your approach was tested for it.

References

Baumhoer, C. A., Dietz, A. J., Kneisel, C. and Kuenzer, C.: Automated Extraction of Antarctic Glacier and Ice Shelf Fronts from Sentinel-1 Imagery Using Deep Learning, Remote Sens., 11(21), 2529, doi:10.3390/rs11212529, 2019.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I. and Rignot, E.: Calving Front Machine (CALFIN): Glacial Termini Dataset and Automated Deep Learning Extraction Method for Greenland, 1972–2019, Cryosphere Discuss., 2020, 1–17, doi:10.5194/tc-2020-231, 2020.

Mohajerani, Y., Wood, M., Velicogna, I. and Rignot, E.: Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study, Remote Sens., 11(74), 1–13, doi:10.3390/rs11010074, 2019.

Xie, Z., Haritashya, U. K., Asari, V. K., Young, B. W., Bishop, M. P. and Kargel, J. S.: GlacierNet: A Deep-Learning Approach for Debris-Covered Glacier Mapping, IEEE Access, 8, 83495–83510, doi:10.1109/ACCESS.2020.2991187, 2020.

Zhang, E., Liu, L. and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13(6), 1729–1741, doi:10.5194/tc-13-1729-2019, 2019.

Interactive comment on The Cryosphere Discuss., https://doi.org/10.5194/tc-2020-310, 2020.