LEARNING DEEP MODELS FROM WEAK LABELS FOR WATER SURFACE SEGMENTATION IN SAR IMAGES

Francesco Asaro, Gianluca Murdaca, Claudio Maria Prati

Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
francesco.asaro@polimi.it

ABSTRACT

Today, one of the biggest challenges faced in the intersection of the Deep Learning (DL) and synthetic aperture RADAR (SAR) domains is the scarcity of precisely annotated datasets suitable for properly training a supervised algorithm. This paper shows that it is possible to successfully exploit weak-labeled data instead of relying on manually annotated labels. In particular, we show how it is possible to train, with state-of-the-art performance, a deep model for the segmentation of water surfaces in SAR images from a weak-labeled dataset. Finally, we present examples of applications of the learned model to the segmentation of inland water bodies and floods.

Index Terms— Deep Learning, SAR, water bodies, floods

1. INTRODUCTION

Precisely annotated datasets are characterized by labels which have a one-to-one correspondence with the input and theoretically zero errors, since they are manually generated by domain-experts. Instead, weak-labeled datasets admit the presence of a significant disturbance in the labels, which can be interpreted in different ways according to the domain. The production of a precisely annotated dataset requires extreme effort and is almost impractical in the Earth Observation (EO) domain, given the scale of its applications. Indeed, the vast majority of publicly available EO datasets are weakly labeled [1], thus generated from data sources at lower spatial, temporal, semantic or thematic resolutions than the input data source, and are produced with semi-automatic procedures. In this paper we use a set of experiments to show how it is possible to successfully train a Deep Convolutional Neural Network (DCNN) in a weak supervised fashion, for the semantic segmentation of water surfaces in SAR images. The possibility of monitoring water surfaces from space has played a fundamental role in the understanding of extreme events such as floods and droughts, as well as in improving the management of water as a natural resource. As climate change continues to trigger extreme weather events, the demand for new tools to monitor water surfaces has never been greater. Thus, the development of better performing algorithms is a key point for transforming EO data into useful insights and then actions. SAR has always been the major player in surface water mapping tasks, thanks to its all-weather capabilities and its physical-based interpretation, which helps address water surfaces through its specific scattering mechanism [2]. Nevertheless, few solutions have been developed in the DL era [3, 4] due to the issues discussed above. Furthermore, these studies focus exclusively on flood mapping applications, neglecting small water bodies, and are trained on a lower resolution ready-to-use versions of the input data, which will be addressed in section 2.1. This paper is structured as follows: in section 2 we describe the dataset, in section 3 we introduce the learning problem, in section 4 we present the experiments and in section 5 we provide the validation and test results.

2. DATASET

In this section we introduce the input data and label sources, as well as the processing steps carried out to assemble the dataset. We also provide details on the statistics of the training, validation and test sets.

2.1. Input data

Sentinel-1 (S-1) mission is the major free source of SAR data. It has a global spatial coverage with a 6-days (Europe) to 12-days (World) revisit time and up to 1 day (Europe) temporal coverage, intended as the time span between two consecutive observations in different SAR geometry. S-1 background operational mode TOPSAR-IW acquires 22×5 meters spatial resolution single look complex (SLC) data. In this work, we build our dataset from full-resolution SLC, exploiting σ_0 calibrated intensity in both VV and VH polarization, rather than resorting to multi-looked ground range detected (GRD) data as is done for the main publicly available dataset [1] and in other works [3, 4]. The raw SLC are processed with a standard pipeline using SNAP, which performs debursting, merging, calibration and geocoding. σ_0 intensities are geocoded to UTM map reference at a sampling space of 10 meters using nearest neighbour in order to not alter the radiometry of the data.

2.2. Labels

The weak labels are extracted from 2015 Copernicus' Water and Wetness High Resolution Layer, which is a 10 meters spatial resolution product. It is generated with a semi-automatic procedure from calibrated NDVI obtained from different high resolution optical satellites [5]. This product has a thematic resolution of 5 classes, permanent and temporary water bodies, permanent and temporary wetlands, and sea surfaces. Refer to table 1 for the id of each of the five classes.

class-id	*class semantic*
1	permanent water bodies
2	temporary water bodies
3	permanent wetlands
4	temporary wetlands
255	sea surface

Table 1. Classes in the Water and Wetness High Resolution Layer.

Being a multitemporal aggregate, this product only depicts the average condition of water surfaces. From the semantic point of view it maps water bodies down to a width of 30 meters, ignoring the smallest one. Thus, it can be considered a weak label due to both temporal and semantic flaws.

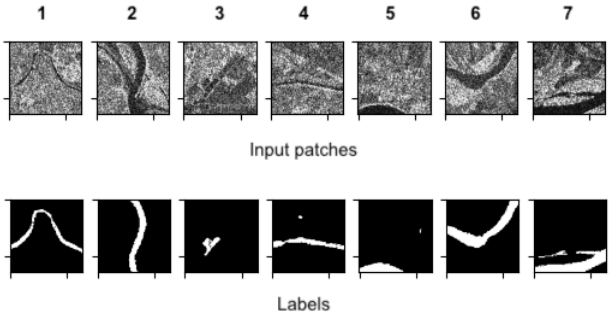

Fig. 1. Sample input patches (VH) and labels.

Indeed, in figure 1 it is possible to observe that the labels do not match the water surfaces in the input patches exactly and that the minor river branch in patch 2 is not annotated in its label.

2.3. Train-test-validation sets and statistics

In EO-related researches, spatial auto-correlation phenomena often hampers validation of supervised algorithm [6]. In order to avoid this mistake, we build our training set by extracting 128×128 patches from a single S-1 full-swath image over the Low Countries (Netherlands, Belgium), while validation and test set patches are sampled in time and space from two different areas of a full-swath stack spanning the Padana Plains (Italy) area at a one year temporal horizon. These two regions are both rich in water bodies but are characterized by significant morphological differences. Multitemporality is exploited in order to test robustness against seasonal variability. The validation and test sets are generated by randomly sampling the available patches respectively from a 70% - 30% split of the Padana Plains area. The patches are not sampled blindly from the two splits, but from subsets of patches with the same positive class frequency, in order to enforce the same global positive class frequency in both the validation and test sets, as shown in table 2.

set	*# patches*	*positive class frequency*
train	13145	0.065
test	5327	0.024
validation	2280	0.021

Table 2. Datasets statistics.

3. DEEP MODEL

In this section we present the details relative to the implemented architecture and the optimization exploited to learn the segmentation task over S-1 images. In this set of experiments the architecture parameters are fixed and are not subject to an ablation study.

3.1. Architecture

U-Net is one of the most successful models for image semantic segmentation and denoising tasks [7]. Originally developed for applications on medical images, it has been exploited in various domains, including EO [8].

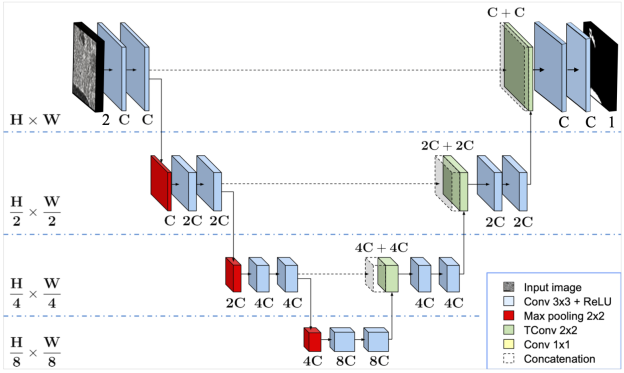

Fig. 2. U-Net architecture.

Its peculiarity is the encoder-decoder structure: the encoder compresses the information by means of a series of pooling operations, allowing the extraction of the most significant features from the input data at different levels of abstraction. Each block of the encoding path is composed by a 2×2 max-pooling followed by two stacked convolutional layers with 3×3 1-strided kernels. Except for the input layer, the encoding block is fed with the feature maps of the previous block and it produces a pooled tensor with doubled features

$$O_E^l = Enc(O_E^{l-1}) \quad (1)$$

where O_E^{l-1} of dimension $H^{l-1} \times W^{l-1} \times C^{l-1}$ is the output of the previous encoding layer and O_E^l are the feature maps extracted at the current level with dimension $\frac{H}{2}^{l-1} \times \frac{W}{2}^{l-1} \times 2 \cdot C^{l-1}$.
The decoder maps the latent representation back to the original spatial resolution by means of a series of upsampling operations. Each decoding block O_D^{l-1} upsamples the features fed by the previous block, applying a 2×2 transposed convolution. The feature maps decoded at l-th level are concatenated with those at the same encoding level in order to help the reconstruction (skip-connections)

$$I_D^l = O_E^l \oplus Upsample(O_D^{l-1}) \quad (2)$$

where O_E^l and O_D^{l-1} are the feature maps extracted by the encoder (E) and by the decoder (D) at the l-th and $(l-1)$-th layers respectively, $\oplus$ represents the concatenation operator. Each convolutional layer is followed by a rectified linear unit (ReLU), a non linearity defined as

$$ReLU(x) = max(0, x). \quad (3)$$

Finally, the features decoded back at the original input resolution are fed to the classification layer, made of two stacked convolutional layers with 1×1 kernels. The raw output logits are then used to evaluate the loss function during training or normalized with a sigmoid during inference. The overall U-Net architecture is depicted in figure 2, it corresponds to the baseline implementation as [7] deprived of one encoding block to accommodate the smaller 128×128 patch size, compared to the 512×512 in [7]; C equals 64.

3.2. Optimization

The architecture's weights are learned by backpropagation training over 100 epochs, using an ADAM optimizer [9] with a learning rate of $1e^{-3}$ and standard Betas. Gradients are computed over mini-batches of 16 patches of size 128×128. The loss function to which is subject the optimization problem is a Binary Cross Entropy (BCE) computed on logits, in order to exploit the sum-log-exp trick

$$L(x, y) = -\sum^{n}[y_n log(\sigma(x_n)) + (1 - y_n) log(1 - \sigma(x_n))] \quad (4)$$

4. EXPERIMENTS

This set of experiments is composed of seven runs aimed at understanding which combination of sub-classes is the best performing (table 1) when used to predict all the sub-classes as a single positive class. Models performance is measured in terms of F_1 score, the harmonic mean of precision and recall. The training and validation learning curve are shown in figure 3. For each run, test F_1 and other performance metrics are computed at the state that produced the highest validation F_1. Performance details are presented in the following section. The experiments are implemented in PyTorch.

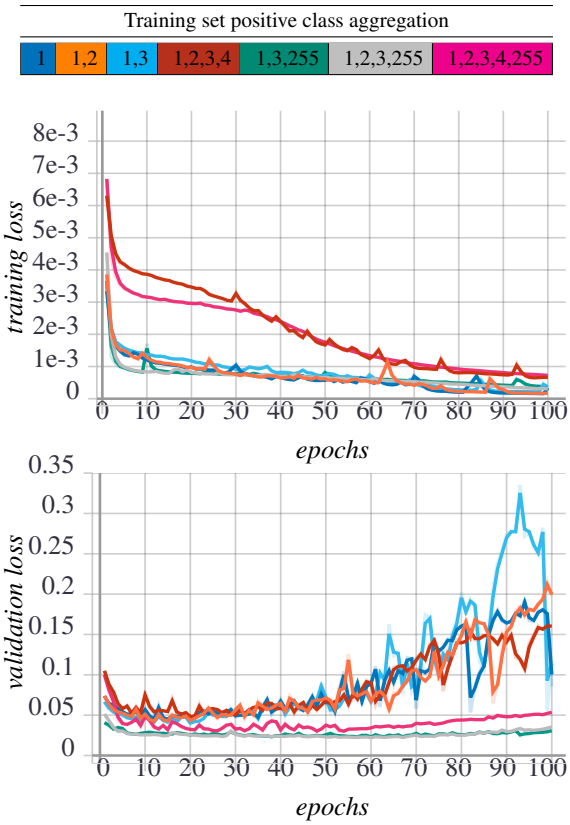

Fig. 3. Training learning curves (top) and validation ones (bottom).

5. RESULTS AND CONCLUSIONS

We have composed a weakly labelled dataset for training a deep architecture to segment water surfaces in SAR images. Table 3 and 4 summarize validation and test performances for the seven different runs. The best test and validation F_1 scores (.87, .86) are achieved by aggregating the sub-classes 1,2,3 and 255. Thus, it is interesting to notice that the subclass 4 (temporary wetlands) acts as a disturbance in the learning phase. The contribution of the class 255 (sea surface) in regularizing the learning also stands out.

class-id	F_1	*IoU*	*precision*	*recall*	*loss*
1	.693	.533	.941	.549	.044
1,2	.695	.533	.901	.566	.045
1,3	.717	.559	.944	.578	.040
1,2,3,4	.767	.622	.844	.703	.051
1,3,255	.858	.751	.937	.791	.022
1,2,3,255	.865	.762	.927	.811	.023
1,2,3,4,255	.857	.751	.872	.843	.030

Table 3. Performances on the validation set at the epoch that produced the best F_1 score for each of the seven runs.

These results also point out that is not necessary to perform speckle-filtering pre-processing. Given the encoder-

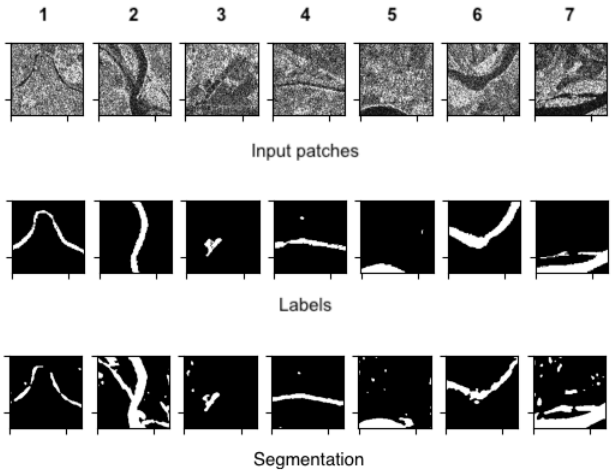

Fig. 4. Test samples segmentation performed by the best model.

class-id	F_1	*IoU*	*precision*	*recall*	*loss*
1	.691	.528	.877	.570	.034
1,2	.663	.496	.843	.630	.040
1,3	.774	.593	.892	.540	.031
1,2,3,4	.748	.597	.782	.716	.055
1,3,255	.863	.767	.896	.841	.020
1,2,3,255	.873	.774	.887	.858	.020
1,2,3,4,255	.855	.746	.821	.891	0.032

Table 4. Performances on the test set at the epoch that produced the best validation F_1 score for each of the seven runs.

decoder structure of U-Net it is plausible to think that segmentation and denoising task are learnt jointly. This speculation will be addressed in future work.

The third row in figure 4 shows the segmentations produced by the best F_1 scoring architecture for some test samples. It is evident that the architecture succeeded in learning to produce detailed segmentations of small water bodies. It is interesting to notice that the architecture is also able to reconstruct the minor river branch not labeled in column 2 and other finer details in the other samples. This fact is representative of the great generalization capabilities of DL architecture. Figure 5 shows an application, always of the best F_1 scoring architecture, for the delineation of a flooded area.

Fig. 5. Nonantola (Italy) flood on 6th Dec 2020

6. CODE AND DATASET

A dockerized inference-ready version of the best architecture is available at `https://github.com/francescoasaro/IGARSS21`. The dataset is available upon request to the authors.

7. REFERENCES

[1] Schmitt, M., Hughes, L.H., Qiu, C. and Zhu, X.X., 2019. SEN12MS–A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. arXiv preprint arXiv:1906.07789.

[2] Henry, J.B., Chastanet, P., Fellah, K. and Desnos, Y.L., 2006. Envisat multi-polarized ASAR data for flood mapping. International Journal of Remote Sensing, 27(10), pp.1921-1929.

[3] Nemni, E., Bullock, J., Belabbes, S. and Bromley, L., 2020. Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. Remote Sensing, 12(16), p.2532.

[4] Pai, M.M., Mehrotra, V., Aiyar, S., Verma, U. and Pai, R.M., 2019, June. Automatic segmentation of river and land in sar images: A deep learning approach. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 15-20). IEEE.

[5] Congedo, L., Sallustio, L., Munafò, M., Ottaviano, M., Tonti, D. and Marchetti, M., 2016. Copernicus high-resolution layers for land cover classification in Italy. Journal of Maps, 12(5), pp.1195-1205.

[6] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N. and Lyapustin, A., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature communications, 11(1), pp.1-11.

[7] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[8] Lattari, F., Gonzalez Leon, B., Asaro, F., Rucci, A., Prati, C. and Matteucci, M., 2019. Deep learning for SAR image despeckling. Remote Sensing, 11(13), p.1532.

[9] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.