Custom state-of-the-art CNN algorithm for ship detection and segmentation

Gianfausto Bottini[a], Marco Corsi[a], Filippo C. Daffinà[a], Simone Tilia[a], Torbjorn Stahl[a], Dino Quattrociocchi[a]
[a]e-Geos S.p.a., via Tiburtina 965, Roma, Italy

ABSTRACT

The use of Deep Neural Networks (DNN) for the detection and segmentation of specific objects from satellite images is a topic of great interest today. This paper presents the results of the application of a CNN (Convolutional Neural Network) state-of-the-art algorithm, Mask R-CNN [4], for the detection and segmentation of ships from Very High Resolution (VHR) and High Resolution (HR) optical images starting from open-source and proprietary annotated VHR data-sets. After having tried several CNN models as Unet and SegNet, the result is that the Mask-R-CNN shows better results both for the detection of ships in the open sea and for detecting ships near the coasts or in case of difficult conditions such as presence of clouds and waves. We are improving the Mask-R-CNN method's detection capability by extending the training dataset using Data Augmentation rotating, flipping, cropping and zooming out images to enlarge the training dataset both on VHR and on HR.

Keywords: Deep Learning, Mask R-CNN, SegNet, Unet, Data Augmentation, Vessel Segmentation, Maritime Domain Awareness

1. INTRODUCTION

The increasing availability of satellite imagery by a heterogeneous set of sensors, results in the strong need to automatize the processing of the huge amount of input raw data, selecting the fastest and most accurate approach. In this context it is very difficult to find a unique solution able to maintain the same level of performance and accuracy over VHR images as well as over HR images. The adoption of already existing Deep Neural Networks, opportunely trained with specific training datasets, is showing the best results over the variety of data provided as input. The aim of the presented activity is to be able to automatically segment every type of vessel from optical images from different satellites using both VHR (Very High Resolution) and HR (High Resolution) sensors.

The original reference dataset on which the deep learning algorithms were driven consists of about 200,000 VHR 768x768 pixels images, 1.5 m resolution, containing carefully labeled ships of every type, and with a test set of about 15,000 specimens to which add more samples from other datasets. This dataset is one of the few freely available regarding ship segmentation. Having assessed the execution of three different existing CNN (SegNet [2], U-Net [3] and [4] Mask R-CNN) at the end Mask R-CNN was selected due to its accuracy boost in case of cloud, waves and coasts in ship detection and segmentation.

The dataset has been expanded with custom samples of HR sentinel2 images carefully labeled by our mapping team and ad hoc data augmentation by means of zoom out technique and the model target is recognizing the shape, the color and the background of the vessels despite the image sensor, HR or VHR.

The final workflow of the software consists of downloading an image, which can be from Sentinel2, SPOT or BlackSky, making predictions on it, detecting and segmenting vessels and producing a digital report containing several information for each of them like:

- length and width of the vessel and of the bounding box around the vessel
- Orientation

- Vessel area in square meters
- coordinates in longitude, latitude
- Classification of the ship in four categories A-B-C-D, depending on its size
- The date of acquisition

The results of the algorithm are stored in a KML file along with the final GeoTiff clip containing the predictions.

2. METODOLOGY AND RESULTS

As already mentioned in the previous section, 3 CNN models have been considered and implemented. SegNet [2] and Unet [3] are considered as the standard architectures of Convolutional Neural Networks with regard to the classification of images, i.e. class recognition plus object segmentation. Both SegNet and Unet are fully convolutional neural networks and provide pixel wise classification that for our target, from a visual analysis, are reaching satisfactory results in favorable conditions.

Mask R-CNN is a regional Convolutional Neural Network at the state-of-the-art and reaches very high accuracy both on training and validation sets and, from a visual analysis on the test set, in comparison with the previous two methods it shows a significant improvement in unfavorable conditions (clouds, coasts, vessels close to each other etc.).

Approximately 90,000 images have been adopted as training set along with some custom HR flagged samples, divided into training and validation sets in the 80-20 proportion and using as test set samples coming from GeoEye, SPOT, DOTA, Black Sky and Sentinel2.

Regarding HR case, a synthetic dataset has been created degrading VHR images and masks by means of Pyramid reducing, an algorithm often used in classic Computer Vision tasks taking resolution pixels from 1.5 m to ~10m, similar to Sentinel2.

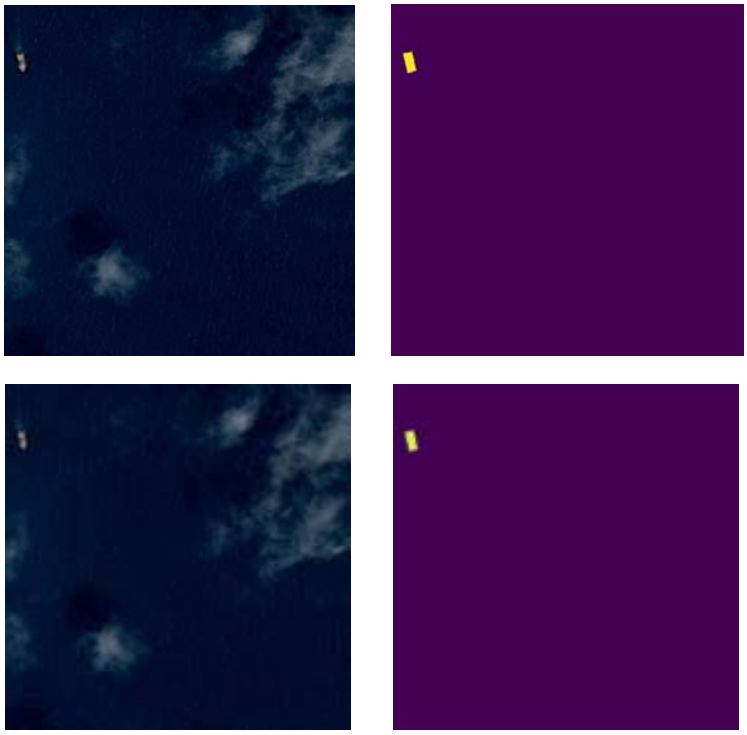

Figure 1. Top-left: VHR image, Top-right: VHR mask, Bottom-left: Synthetic HR image, Bottom-right: Synthetic mask

For Segnet [2] and Unet [3] the function to be minimized is the Binary Cross-entropy that outputs the probability of each pixel being part of class 0 and class 1. The Adam Optimization Algorithm was selected for optimizing the results, instead of using the classic Stochastic Gradient Descent algorithm, and models were trained from scratch with no Transfer Learning Technique

The Mask R-CNN [4] model was initialized with the weights obtained on the MS COCO[6] dataset. For this model, the application of Transfer Learning is used to make the network more efficient in less time. The model minimizes the loss on the class, bounding box and segmentation, and is therefore a linear combination of the 3; The optimizer is the Stochastic Gradient Descent with appropriate values for momentum and learning rate parameters.

Data Augmentation has been an important window of the final model improving and increasing the dataset by several thousands of samples such that the model is customized over the target to segment images coming from any source.

The metrics used to evaluate the accuracy for all three methods for the VHR case is the harmonic-F1-score with the following results:

Table 1. Harmonic-F1-score.

Model	F1 score train	F1 score validation
SegNet	0.7	0.6
Unet	0.79	0.69
Mask R-CNN	0.84	0.75

The confusion matrix is used to have a general idea about the number of ships detected in a validation set of about 15,000 samples as follows:

Table 2. Confusion matrix.

	Ship forecasting	Non-ship forecasting
Ship real	80 %	1%
Non-ship real	1.2%	70%

where the fourth cell is considered as the percentage of cases in which the algorithm predicts just the exact number of vessels.

Once the training is over, the detailed software online workflow is as follows:

- The image is downloaded from its source and is cut into sub-tiles, the size of the sub-titles is a parameter given by the operator based on the size and the source type.
- Each sub-tile is processed by the neural network, previously trained, creating a map of vessels identified within the sub-tile.
- Each of the identified ships is fitted into an ellipse that best characterizes it. Usually, the input images are in the GeoTiff format helping us to shift pixels into real measures like length, width, area in square meters and providing us the coordinates in longitude and latitude.
- Each sub-tile is then merged with the others into a final prediction and a final digital report containing all this information.

The final goal is to have a continuous Maritime Awareness picture with precise and reliable information in near real time. The processing time of the algorithm is less than 1 minute for a Sentinel2 image of 10,000 pixels by 10,000 pixels.

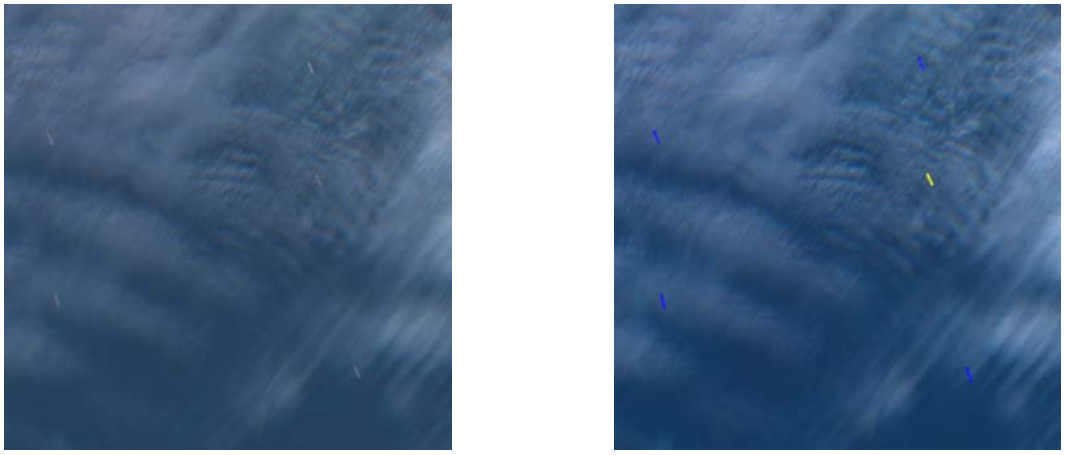

Figure 2. On the left the raw image (Sentinel2), on the right the forecasted image

In Figure 2 above is illustrated an example starting from raw data, a Sentinel2 image in open sea close to the Khark island. The image is cut and processed by means of the Neural Network weights, and each sub-tile has its prediction that then are put together in a final prediction over the entire image as it is shown on the right; an example of the final digital report is given below. The result is stored in a KML file, which opened on the Google Earth portal gives the following information:

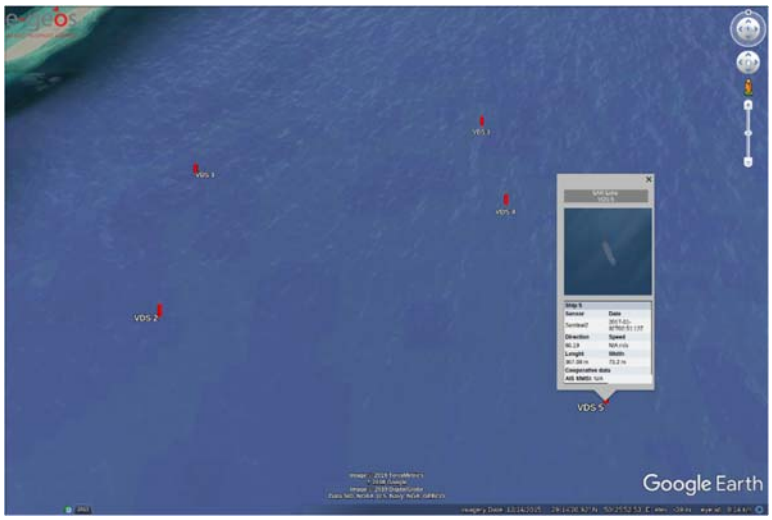

Figure 3. The digital report consists of a KML file that can be displayed in Google Earth with exact vessels position

Figure 4 gives some examples of forecasting over images coming from several different VHR sensors such as SPOT, DOTA, Google Earth.

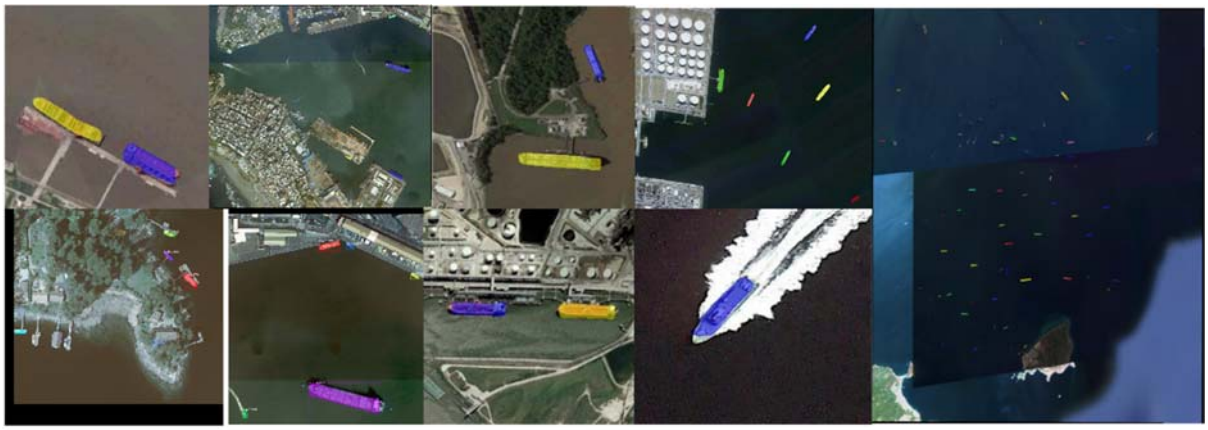

Figure 4. Example of vessels segmentation on Satellite acquisitions coming from several different VHR Sensors

3. CONCLUSIONS

Results have been very promising and comforting and have been showed in different presentations and workshops including the "Ital-IA" [1] national conference about Artificial Intelligence for Space but the project flow is not over. Preliminary application and operational validation of the selected method will be performed during the next maritime trials in the context of H2020 project MARISA [6]. Our future challenge will be to enlarge the HR dataset with synthetic datasets by means of Generative Adversarial Networks, as well as by degrading VHR images to HR by means of custom filters, to be able to entirely drop the manual work on HR. The work of generating the synthetic dataset has already begun with promising results. Although the Mask R-CNN[4] weights trained on the VHR dataset give fairly acceptable predictions also on the HR dataset, drawing a new network on a flattened HR dataset (not existing today) would give much better results for these datasets. The idea is to follow the concept of particular generative neural networks called CycleGAN [5] that transfer style and ownership from one image dataset to another. The goal is to create datasets of HR, simil-Sentinel2 and Landsat8-like images, starting from VHR datasets and degrading images, leading to labeled HR images. A CycleGAN [5] consists of three networks, two generative and one discriminatory. The first to convert a VHR image into an HR, the second to convert an HR into a VHR and the third to identify whether the image is HR or generated. We want to create a continuous flow of generation and discrimination of satellite image data of all kinds, based on the latest models of artificial intelligence, and to be able to generate, govern, recognize and label every type of image coming from every type of satellite, which is one of the goals of every company involved in the Earth Observation field.

REFERENCES

[1] Bottini, G., Corsi, M., Daffinà, F. and Quattrociocchi, D., "Analysis of Deep Learning models for Ship Detection on VHR and HR optical satellite images", Ital-IA (2019).

[2] Badrinarayanan, V., Handa, A. and Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling", Badrinarayanan (2015).

[3] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation", Ronneberger (2015).

[4] Ke, K., Gkioxari, G., Dollàr, P. and Girshick, R.,. "Mask R-CNN", He (2017).

[5] Zhu, J. Y., Park, T., Isola, P. and Efros, A. A., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", Jun-Yan Zhu (2018).

[6] COCO dataset, "COCO – Common Objects in Context", <http://cocodataset.org/#home> (2019).

[7] H2020 MARISA project website, "MARISA - Maritime Integrated Surveillance Awareness", <http://www.marisaproject.eu> (2019).