ON PERTURBATION-BASED XAI FOR FLOOD DETECTION FROM SAR IMAGES

Anastasia Schlegel, Ronny Hänsch

German Aerospace Center (DLR), Microwaves and Radar Institute, Germany

ABSTRACT

Excelling in various image analysis tasks, machine learning (ML) models and especially deep convolutional networks (ConvNets) have become a cornerstone in the Remote Sensing community. However, their complexity makes their decision-making process opaque, rendering deep ConvNets as black box models. To address this issue, "Explainable AI" (XAI) methods have been proposed that aim to provide insights into the rationale behind ML generated predictions. Amongst them, perturbation-based techniques monitor changes in the prediction related to local distortions of the input. Thereby the relative importance of the altered input area for the prediction is determined that serves as an explanation for the network's prediction. In the context of flood detection from SAR images, we investigate the impact of different parameter settings on the relevance estimation and thus on the explanation. The experimental results indicate a strong parameter dependence yielding ambiguous and partly contradicting explanations.

Index Terms— Explainable Machine Learning, Occlusion, Synthetic Aperture Radar (SAR), Flood Detection

1. INTRODUCTION

Over the last few years, the use of Machine Learning (ML) has steadily increased in the Remote Sensing (RS) and Earth Observation (EO) community. Particularly Deep Neural Networks (DNNs) have been established as the model of choice for various image analysis tasks. However, the strong performance of DNNs is linked to their complex structure of millions of interdependent variables inside a nested nonlinear function. Although all relevant parts within a neural network are in principle accessible, their sheer number, extreme connectivity, and interdependence make the exact functional relationships nontransparent. This causes a lack of explainability and interpretability of the actual function the model has learned and the results it produces, giving DNNs the reputation of being black box models.

The umbrella term "explainable AI" (XAI) refers to approaches with the aim of opening black box models and understanding their predictions. Rather than creating so-called white-box models that are inherently interpretable, XAI methods provide explanations to already trained models. Such post-hoc techniques are commonly categorized by their specificity. While gradient-based approaches [1] consider how the model or parts of it contribute to the the prediction by leveraging gradients computed at individual instances, perturbation-based approaches [2] observe changes in the prediction related to local distortions of the input.

While both groups have merit, perturbation-based methods have the advantage of simplicity and being model-agnostic, i.e. they can be applied to any kind of machine learning model that is treated as a black box. They compute the attribution, i.e. the relevance, of an input feature by masking or altering it and monitoring the difference in prediction for the new input to the original input. From an implementation perspective, perturbation-based methods only require repeated forward passes, without the need to understand the model's inner workings. Comparably simple is the interpretation of their explanations, considering that the computed relevance maps are directly linked to the model's output variables. However, with an increasing number of test features, perturbation-based methods come with the cost of a certain computational load. As more critical we consider the strong dependence of the relevance estimation on the method's parameters, i.e. the size of the altered patch as well as the value replacing the original signal. Given that input patches can be modified arbitrarily, we assume a certain ambiguity in the explanations perturbation-based methods generate.

This paper presents a survey of perturbation methods used to generate explanations in the context of flood detection from Synthetic Aperture Radar (SAR) images. In the context of natural hazards, in particular flood events, explainability is of utmost importance since model predictions are potentially used by responders to organize help or rescue efforts, perform damage assessment, monitor the progress of the flood, etc.. A mere prediction of which areas are flooded is often not sufficient but explanations for why the model arrived at a certain decision are required as well. To focus on the analysis of explainability, we employ a rather standard DNN, i.e. a U-Net [3], which is trained and evaluated on the Sen1Floods11 [4] dataset. We evaluate different parameters of a perturbation-based method, in particular the size and type of the occlusion, and how they influence the estimated relevance maps.

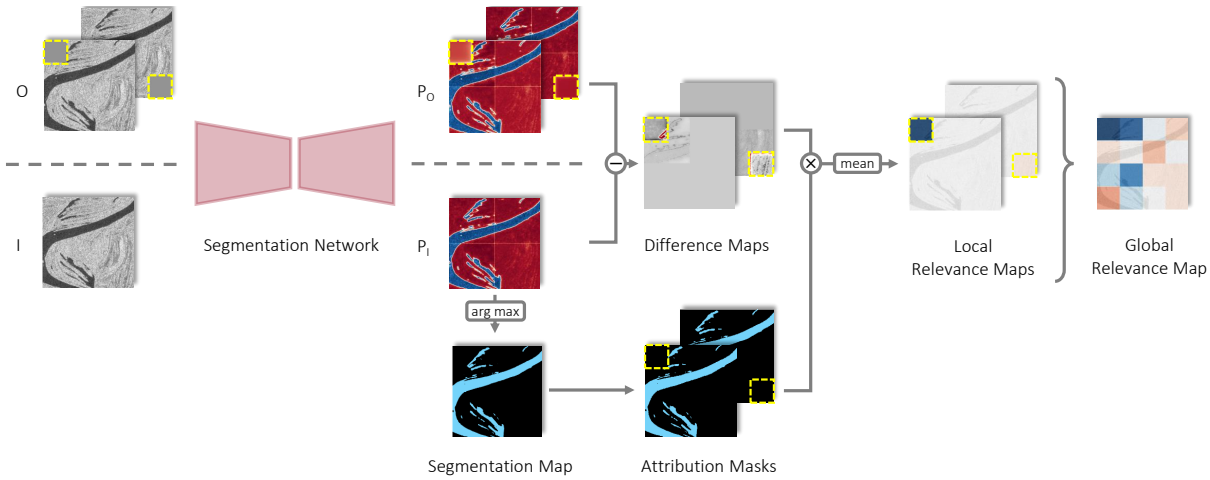

Fig. 1. The occlusion method determines the relevance of individual image patches for the prediction of water areas in SAR images. Within the probability maps P_I and P_O, blue areas represent high probability values for the presence of class water, while low probability values are shown in red. In the difference map, red is associated with a prediction drop, black with an improved prediction, and gray with no change in prediction for the occluded image with respect to the original image.

2. METHOD

One approach to answer the question whether a model prediction is based on the object itself or the surrounding context, is the occlusion method [5] which monitors the change of the models output if an input image with a locally distorted region is used. A decrease in the classification probability indicates the relative importance of the occluded patch for the prediction, while a probability increase is attributed to a disturbing influence of the occluded patch on prediction.

We extend this approach from image classification, i.e. assigning one label for a whole image, to semantic segmentation, i.e. every pixel is assigned a semantic label, and apply it to the use case of flood detection in SAR data. Fig. 1 provides an overview of the relevance estimation of individual image patches using the occlusion method for the prediction of water areas in SAR images. Given are input images I and O, where I shows the original image and O an occluded version. The set of occluded images O is created from I by shifting an occlusion window over the original image, thereby blocking out parts of the original signal. Forwarding I and O individually through the segmentation network, it outputs the water class probability on a pixel level, i.e. the probabilistic maps P_I and P_O. For each I and O pair, we compute the relevance score R_O of an image area from the probability difference $P_I - P_O$.

To attribute the relevance of single image patches to the prediction of a certain region, e.g. a water segment, we create attribution masks from the segmentation output of the original image I that combine water pixels to larger segments. Averaging the difference values over the water segments results in a relevance score of the respective patch for the prediction of water, stored in a local relevance map. To avoid the bias coming from large difference values within the occluded image part itself, this region is ignored in the attribution masks in the relevance estimation. The individual relevance scores are combined in a relevance map of the same size as the original image I. Patches with positive relevance scores (blue) are considered important for the prediction, while patches with negative relevance scores (red) indicate a disturbing influence.

3. EXPERIMENTAL SETUP

We analyze the occlusion method for semantic segmentation and employ a U-Net, which we train and evaluate on the Sen1Floods11 [4] dataset. The dataset consists of 446 hand-labeled Sentinel-1 images in VV and VH polarization collected from flood events across 11 countries. According to [4], the 512×512 images are split into four 256×256 patches in the inference stage and are stitched back together later. This means, as illustrated by the difference maps shown in Fig. 1, that the distortion only affects the prediction within a 256×256 patch and does not affect the prediction of water areas inside the remaining patches that together form the original 512×512 image.

We investigate the influence of two occlusion parameters, i.e. the size of the occluding patch and the occlusion type, on the relevance estimation and thus on the explanations quality. We systematically cover image parts with patches of size 8×8, 32×32 and 128×128 to observe how the relevance scores change for smaller and larger occlusion windows. To understand the influence of the occlusion value, we consider the following five cases:

(A) The patch stores the channel-wise mean of $\bar{x}_{VV} = 0.84$ and $\bar{x}_{VH} = 1.01$ as a single, constant value.

(B) Each pixel value inside a patch is randomly sampled from a normal distribution where mean and standard deviation are estimated from pixels belonging to the water class. For each patch, ten occlusion windows are created by drawing new samples leading to ten relevance maps which are then averaged.

(C) Each pixel value in the patch is randomly sampled from a normal distribution where mean and standard deviation are estimated from pixels surrounding the patch. As above, ten occlusion windows are created and their relevance maps averaged.

(D) Using ten different seeds, the pixels inside a patch are randomly shuffled. For each seed a relevance map is produced which is then averaged.

(E) The whole patch is replaced by another patch extracted from a different image. Using ten external patches, the individual relevance maps are combined to an average.

4. RESULTS AND DISCUSSION

Figure 2 summarizes the relevance maps obtained for the different experimental settings. The size of the occlusion patch indicates a trade-off between the resolution of the relevance map and the influence of individual patches on the prediction. As it can be expected, the relevance score decreases with the size of the occlusion patch. Covering larger image parts with patches that carry different information than the original image, naturally has a greater impact on the prediction compared to when smaller image parts are distorted. Structures that are relevant for the prediction might not only be distorted but entirely removed. An occlusion window of size 128×128 for instance already covers a quarter of the image. While larger occlusion patches yield higher relevance scores, and thus in turn have a greater impact on the prediction, the relevance maps have a coarse resolution. Their interpretation is not straightforward, as the individual patches mostly cover both water and non-water areas. Combined in a single relevance score the contributions of both classes are averaged out and cannot be distinguished from each other anymore. Although an occluded patch is estimated to have a contributing or disturbing influence on the prediction, it is unclear what exactly the influence is attributed to. Thus, such coarse relevance maps have little informative value and do not add to explainability. On the other hand, relevance maps estimated from smaller occlusion windows yield more detailed relevance maps that follow the underlying image structures and allow to attribute relevance values to specific image features. Although this improves explainability, the influence small image patches have on the prediction is particularly small, such that they might be neglected by the network when forming the prediction.

The occlusion value primarily determines the relevance scores of individual image patches and therefore also the overall appearance of the relevance map. Whether water, non-water or the bank region separating the two classes have a positive or negative contribution to the prediction, depends on how well the occlusion value represents the original signal. The prediction of water areas for instance deteriorates in Cases A and E when water bodies are occluded. Given that both, the mean of the test image as well as the external patches are dominated by background pixels, water bodies cannot be represented well enough leading to worse predictions. On the other hand, occluding non-water areas by such patches barely has any effect. However, given that there is no variance between the pixels inside an occlusion patch in Case A, the patch appears rather smooth to the network yielding slightly better predictions when applied on non-water areas. The pixel variance and randomness that is present in Case F stands in contrast to that. While in Cases A and E the patch rather represents the background, in Case B the pixels inside a patch are randomly sampled values from the normal distribution of water pixels. Thus, the prediction is barely affected when covering water bodies by such patches, but it deteriorates in case of the non-water areas. We observe more pronounced relevance scores in the bank region separating water from non-water pixels. Occluding border pixels by water-like patches in a way enlarges the water bodies and thereby increases the probability for actual water pixels being classified correctly. A similar effect can be observed in Cases C and D. While in Case C statistics of the surrounding area are introduced as occlusion values, in Case D the statistics inside the patch are kept. However, through shuffling, image features are disrupted that might be important for the classification of water areas. In both cases the bank area separating water from non-water pixels is distorted, such that water-like pixels appear in the non-water area and vice versa. This variance within and between occlusion patches explains that some pixels in the bank area have a favoring and some a disturbing influence on the prediction. Overall, the results indicate the network's ability to correctly classify the image pixels. For the estimation of relevance scores, a clear separation of the two classes appears to be more important than preserving closed shapes.

5. CONCLUDING REMARKS

With explainable AI a wide range of methods have been proposed that aim to provide an understanding of ML models and their predictions. Perturbation-based XAI approaches monitor changes in the prediction linked to modifications of the input and estimate the relevance of the altered input for the prediction. They are particularly popular due to their simple implementation and allegedly straightforward interpretation of the explanation. To this aim, we investigate several variations of the occlusion method using different parameter choices in the context of flood segmentation in SAR images.

Our experimental results indicate a strong influence of the method's parameters on the relevance estimation. While the explanations obtained from the occlusion method demon-

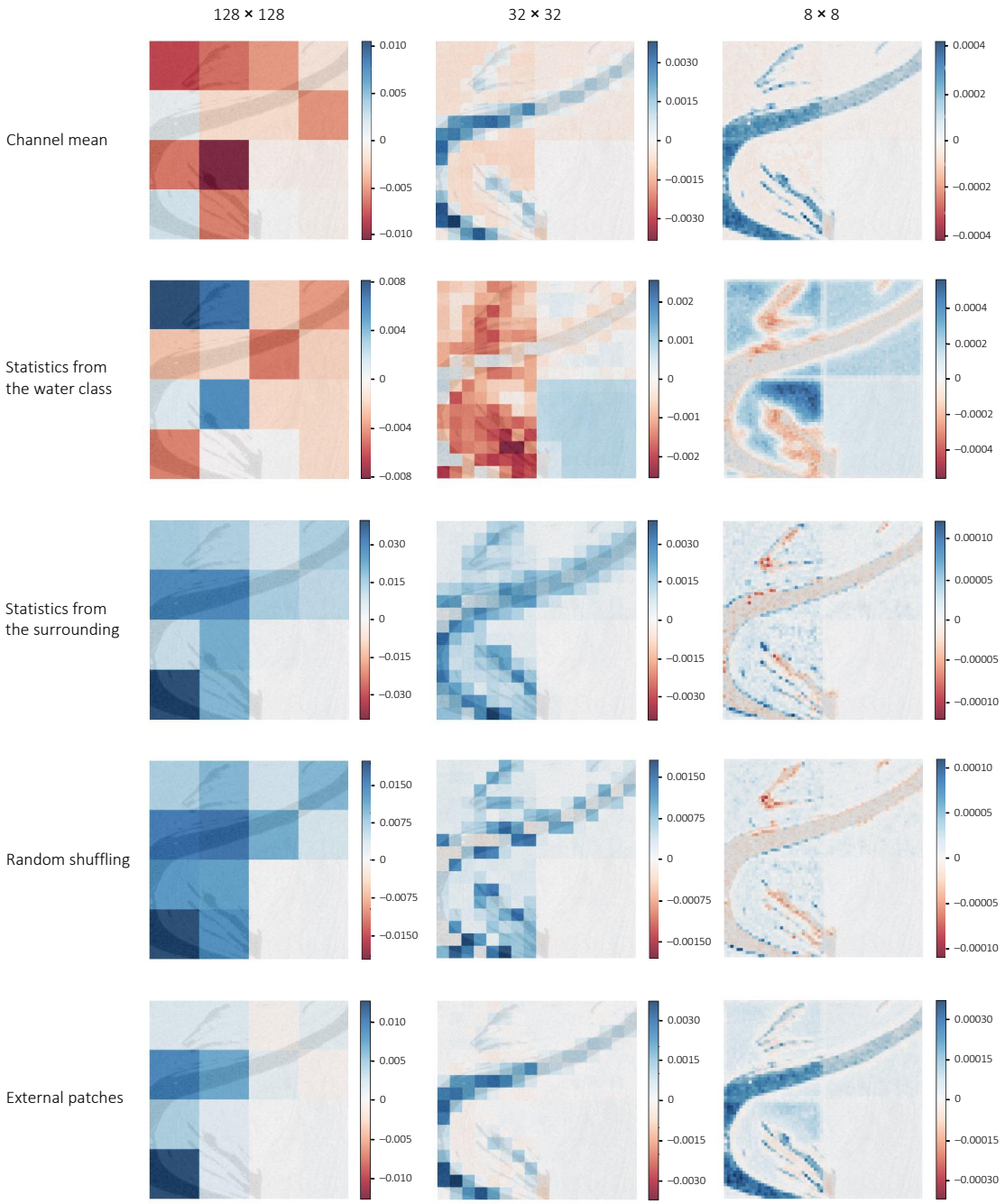

Fig. 2. Influence of occlusion parameters on the relevance of image patches for the prediction of water areas in SAR images. Five occlusion types (channel-wise mean, statistics of the surrounding area, statistics of class water, random shuffling, and external patches) are arranged in rows for three patch sizes (128×128, 32×32, 8×8) arranged column-wise. Blue patches are attributed to positive relevance scores and are considered important for the prediction of the water class. Red patches on the other hand stand for negative relevance scores and indicate a disturbing influence on the prediction.

strate the network's ability to perform the downstream segmentation task, the estimated relevance maps are not consistent with changed parameter settings. Besides the ambiguity of explanations, another drawback of the method is its limitation to only indicate whether an input patch has a contributing or disturbing influence on the prediction. It does not allow to conclude whether the prediction changes because the original values are missing or because new information is inserted.

6. REFERENCES

[1] Marco Ancona, Enea Ceolini, A. Cengiz Öztireli, and Markus H. Gross, "A unified view of gradient-based attribution methods for deep neural networks," *CoRR*, vol. abs/1711.06104, 2017.

[2] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognition Letters*, vol. 150, pp. 228–234, 2021.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[4] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg, "Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[5] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.