ON THE RELEVANCE OF SAR AND OPTICAL MODALITIES IN DEEP LEARNING-BASED DATA FUSION

Jakob Gawlikowski
DW-DAI JE - Data Analysis and Intelligence, DLR
jakob.gawlikowski@dlr.de

Nina Maria Gottschling
MF-DAS OP - EO Data Science, DLR
nina-maria.gottschling@dlr.de

ABSTRACT

When preparing SAR-optical fusion datasets, cloudy samples are often removed from the optical component if they do not contain any information for the prediction task. Although optical data contains more information that is easier to extract and SAR data is noisier, the latter is less affected by changes in the location or illumination and is not obscured by cloud coverage. By removing clouds from the dataset, the realistic situation of cloud coverage is withheld from the network during training and SAR data has less influence on the prediction than when training with cloudy data. In this work, we show on publicly available pre-trained networks and two remote sensing datasets that the effort to filter and correct clouds might not be needed. In contrast, the results of self-trained ResNet18 networks indicate that having cloudy examples in the dataset might lead to a more informative feature extraction from the SAR modality. This leads to networks that utilize the SAR modality comparatively more for predictions, which we show by an increased relevance of the SAR modality. Moreover, such networks obtain improved accuracy, not only on cloudy test samples but potentially also on clear test data.[1]

1 INTRODUCTION

With over 50% of the earth's land surface covered by clouds at all times (King et al., 2013), making a time continuous monitoring of the planet with optical sensors is impossible. While this fact is one of the main motivations to combine the complementary modalities of SAR and optical data (Mahyoub et al., 2019), the optical data in proposed datasets is often cleaned and post-processed to avoid atmospheric distortions and clouds in the optical images (Schmitt et al., 2019; Sumbul et al., 2021). Due to efficiency, accuracy, and scalability to ever-increasing amounts of such remotely sensed datasets, deep learning (DL) based methods have been used for different analysis and prediction tasks (Zhu et al., 2017). In particular, DL-based methods are usually not trained with cloudy samples, and it was shown how this could lead to false predictions with high confidence values when confronted with clouds at test time (Gawlikowski et al., 2022). This is caused by a shift in the data distribution between training and testing data, and data-driven approaches are known to be sensitive to such distribution shifts. SAR-optical data fusion is widely applied in remote sensing and is motivated by complementary properties. While the optical modality is richer in information and more accessible to learn from for DL approaches, the SAR modality brings the advantage of robustness to changes in the illumination, clouds, and atmospheric disturbances (Mahyoub et al., 2019).

For the SAR-optical land cover classification task, multiple large datasets are available for free, including pre-trained networks (Schmitt et al., 2019; Sumbul et al., 2021). In general, these datasets are explicitly cleared from samples where the optical part is affected by atmospheric distortions or insufficient illumination. However, multiple dataset extensions allow applicants to match the original dataset with labeled cloudy examples, as done by (Gawlikowski et al., 2022). The same authors further explain the different effects of clouds on a DL-based classification pipeline and show that neural networks that have not been trained with cloudy optical inputs can give wrong predictions while stating high confidence in the prediction.

While explainability approaches for Machine Learning and DL predictors are already widely applied in uni-modal remote sensing, there is still room for improvement regarding quantifying the

[1] https://github.com/JakobCode/SAROpticalShap

relevance of individual data sources. Multiple approaches motivated by concepts from game theory, such as Shapley values, have been introduced (Gat et al., 2021; Lundberg & Lee, 2017; Parcalabescu & Frank, 2022). These approaches infer the relevance of individual modalities and groups of modalities by the effect of changes in the predictions caused by modifications in other modalities. In general, this procedure is computationally expensive, as the marginalization of individual modalities is approximated via sampling and multiple forward passes. (Hu et al., 2022) introduced SHAPE relevance scores, which have been shown to be an efficient alternative to the more expensive existing approaches. Drawing inspiration from game theory, they also achieve results comparable to more costly methods by omitting modalities and setting the corresponding inputs to uninformative baseline values.

In this work, we investigate how the absence of cloudy samples in the training data affects the fusion of optical and SAR modalities. We do this by closely examining each modality's individual relevance values from different state-of-the-art approaches. We can show how clouds in the training and testing affect the relevance of individual modalities and draw a direct link to model performance. We underline this link by empirical results on a ResNet18 network trained with and without cloudy samples in the training data.

Our contribution in this paper is the following: 1) We investigate the relevance - using SHAPE scores (Hu et al., 2022)- of the individual modalities in SAR-optical data fusion for land cover classification. 2) We show empirically how cloudy samples in the training lead to a more prosperous feature extraction, more balanced modality relevance scores, and more relevance and better performance under the appearance of clouds. 3) We further present empirical results that indicate that the improved information extraction from the SAR modality could also positively affect the classification performance under clear data.

2 METHODOLOGY

As we are interested in the relevance of the whole dataset, we utilize the SHAPE (Hu et al., 2022) approach for our investigations.We follow the notation of (Hu et al., 2022) and denote f as a data fusion neural network, the set of modalities as $\mathcal{M} = \{M_{\text{SAR}}, M_{\text{opt}}\}$, two baseline values 0_{SAR} and 0_{opt} for the two modalities, a model-specific scaling value $Z_f > 0$ and a measure of performance $V_f(\cdot, \cdot)$ (i.e., accuracy), which is computed with the predictions of f based on the given modalities. For the baseline values, we follow (Hu et al., 2022) and replace the input of the corresponding modalities with a zero tensor of the same shape. The SHAPE score is then computed as the Shapley value on the availability of the corresponding data source, in our two-modality case with the modalities M_{SAR} and M_{opt}. For the SAR modality, we obtain

$$S_{\text{SAR};f;\mathcal{D}} := \frac{1}{2Z_f}\left[V_f\left(M_{\text{SAR}}, M_{\text{opt}}\right) - V_f\left(0_{\text{SAR}}, M_{\text{opt}}\right) + V\left(M_{\text{SAR}}, 0_{\text{opt}}\right) - V_f\left(0_{\text{SAR}}, 0_{\text{opt}}\right)\right] , \quad (1)$$

and equivalently for the optical modality. The SHAPE score is bounded between 0 and 1, and a higher value corresponds to a higher relevance. As we want to evaluate the connection between the model performance under different data setups and the modality relevance, we set $Z_f = 1$. (Hu et al., 2022) apply SHAPE to classification tasks, using the accuracy as the performance metric V_f. We extend this procedure and consider the tasks of single-label and multi-label land cover classification. Hence, we apply the accuracy and different versions of the F1 and F2 scores, as described, for example, by (Sumbul et al., 2019).

3 DATA AND EXPERIMENTS

Data SEN12MS (Schmitt et al., 2019) and BigEarthNet-MM (Sumbul et al., 2021) are two widely used datasets when it comes to the fusion of optical and SAR data for land cover classification. For SEN12MS, single- and multi-label classification targets exist; BigEarthNet is stated only as a multi-label classification task. In both cases, the dataset comes with pre-trained models but without clouds and other distortions of the optical data in the training, validation, and test split. However, BigEarthNet also provides additional cloudy samples, and SEN12MS combined with SEN12MSCR contains cloudy images for a subset of the locations in SEN12MS that are used for SAR-optical based cloud-removal (Ebel et al., 2022). For detailed information on the dataset, we reference the

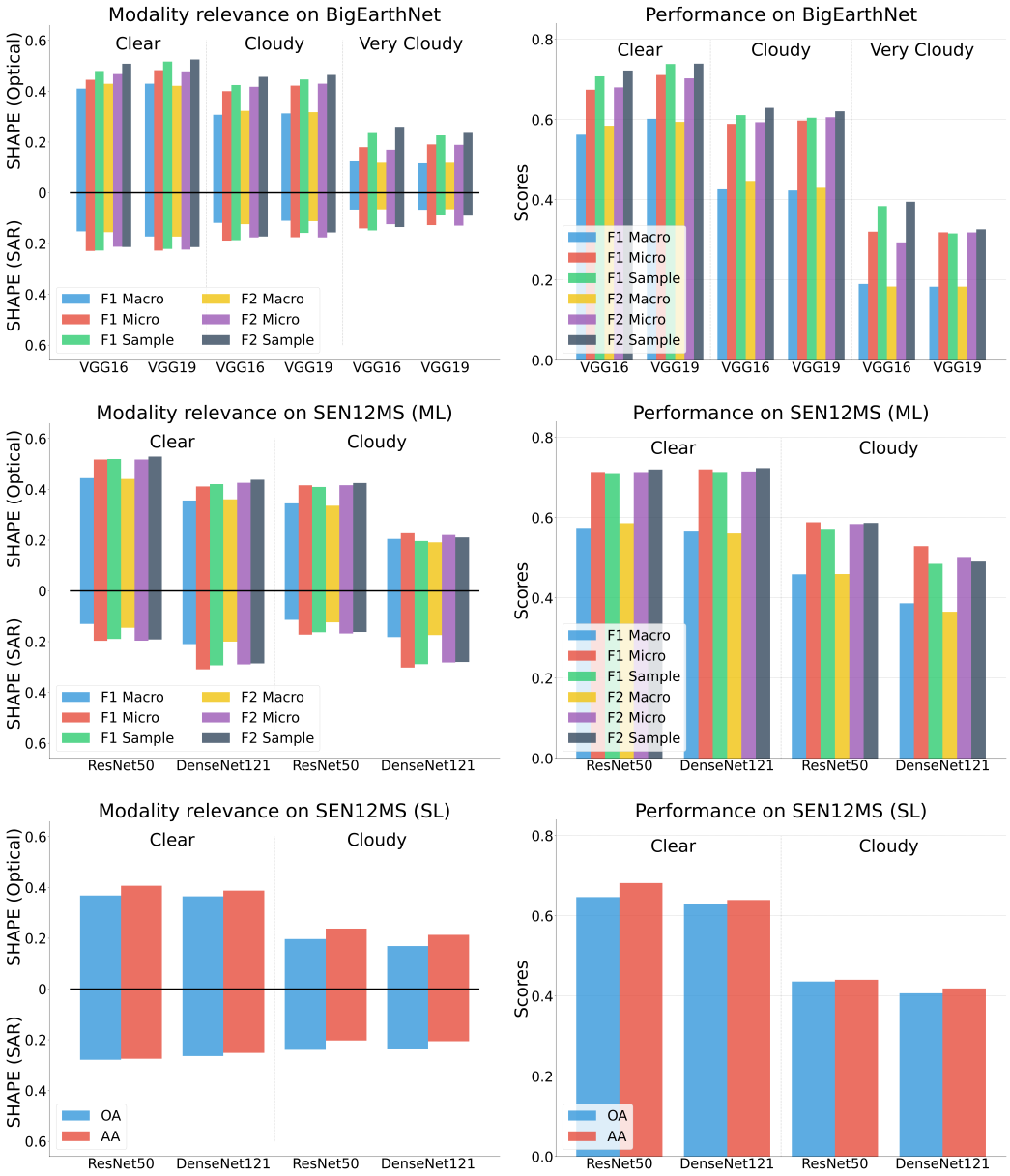

Figure 1: The modality relevance (left) and the classification performance (right) of pre-trained networks on BigEarthNet and SEN12MS with multi-label (ML) and single-label (SL) targets. OA and AA are overall and class-average accuracy. The metrics in the left column represent the choice of V_f.

readers to the original works, (Schmitt et al., 2019; Sumbul et al., 2021), to (Gawlikowski et al., 2022) for further details on the cloud distribution in SEN12MSCR, and to the supplement.

Evaluating pre-trained models We evaluate the pre-trained models available for SEN12MS (ResNet50, DenseNet121 - both for single- and multi-label classification) and BigEarthNet (VGG16 and VGG19) under different data setups. Figure 1 shows the modality relevance and classification performance of these networks. We obtain the same performance for the clear data as in the dataset papers (Schmitt & Wu, 2021; Sumbul et al., 2021). With clouds in the test dataset, the performance drops significantly, most visible for the "very cloudy" subset of BigEarthNet. Both the optical and

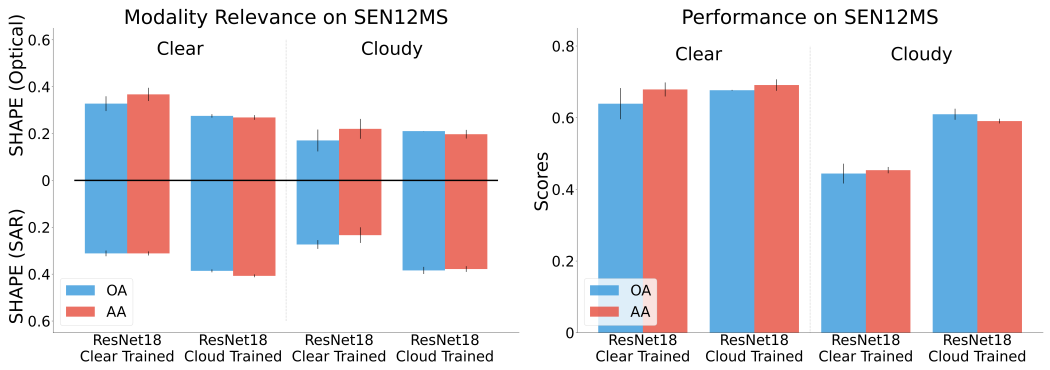

Figure 2: Modality relevance (left) and classification performance (right) of ResNet18 models trained with and without clouds in the training data. The results are based on five runs and show the mean values and standard deviations.

the SAR relevance decrease. For BigEarthNet, the SAR relevance shows a more substantial decrease than in the corresponding evaluations of SEN12MS.

Training own models We utilize the co-registered cloudy-clear samples of SEN12MS and train ResNet18 (He et al., 2016) models on the original and cloudy datasets. The cloudy dataset contains samples with cloud coverage ranging from 0% to 100% with a peak in 90%-100% coverage (Schmitt & Wu, 2021; Gawlikowski et al., 2022). For the cloudy dataset, randomly choose for each sample - if available - whether we load the sample from SEN12MS or SEN12MSCR. We set the sample distribution to 80% probability for SEN12MSCR. We do this as the SEN12MSCR dataset is not fully cloud-covered and represents only a subset of the original clear dataset. For each setup, we train five networks for 25 epochs, use a batch size of 64, and optimize them with the Adam optimizer and PyTorch default parameterization. Figure 2 shows the resulting relevance scores and classification performance values. Comparing the networks trained with and without clouds, one can see that the optical modality is slightly less relevant, while the SAR modality becomes more relevant. Regarding the performance, the networks trained on the cloudy data achieve significantly higher performance on the cloudy test set and a little higher performance on the clear test set.

4 Discussion and Outlook

Discussion The evaluation shows that for the pre-trained networks, the relevance of the optical modality is higher than the relevance of the SAR modality. Regarding the cloud-free training procedure and the richer information and easier access to the optical data, this fits the expectations and findings in other works. Further, under the occurrence of clouds, the model performance and the modality relevance values are lower. Also, the relevance of the SAR modality decreases with the occurrence of clouds, indicating that the SAR modality is not fully explored and mixed with the information from the optical modality at an early stage. We see a higher relevance on the SAR than on the optical component when training with cloudy samples. This indicates that the cooperation among the different data sources is enhanced but that SAR-specific features are also used for (parts of) the classification tasks. This can also be seen in the classification performance, where the performance of the cloudy-trained network is (as expected) better on the cloudy test samples. Interestingly, these models also perform slightly better on the clear test set and with a smaller deviation in the performance. This makes sense when we assume that the network extracts more features from the SAR component, which is less sensitive to regional and illumination changes that appear between the training and the test data. However, as this study is relatively small, further experiments must be investigated to investigate these observations with more repetitions, datasets, and especially more complex models than the used ResNet18. Despite this, it is clearly visible that the cloudy training data did not lead to a clear drop in performance, indicating that the SAR modality is not fully utilized in the clear training setup.

Outloook In the future, we are planning to extend the presented evaluations across a more comprehensive array of setups, data types, and the training of larger and more advanced models. Further, we plan to investigate the development of relevance scores over the training epochs to get more insight into the learning procedure of multi-modal neural networks. Lastly, we aim to extend our investigation to the relevance of individual data points to identify specific strengths and weaknesses of specific data sources, to learn about the qualities of the modalities for the individual classes, and to quantify the capabilities to use the relevance values for the detection of false predictions.

REFERENCES

Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. Sen12ms-cr-ts: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

Jakob Gawlikowski, Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu. Explaining the effects of clouds on remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9976–9986, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016.

Pengbo Hu, Xingyu Li, and Yi Zhou. SHAPE: An unified approach to evaluate the contribution and cooperation of individual modalities. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3064–3070. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/425. URL https://doi.org/10.24963/ijcai.2022/425. Main Track.

Michael D King, Steven Platnick, W Paul Menzel, Steven A Ackerman, and Paul A Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE transactions on geoscience and remote sensing*, 51(7):3826–3852, 2013.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

S Mahyoub, A Fadil, EM Mansour, H Rhinane, and F Al-Nahmi. Fusing of optical and synthetic aperture radar (sar) remote sensing data: A systematic literature review (slr). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:127–138, 2019.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. *arXiv preprint arXiv:2212.08158*, 2022.

Michael Schmitt and Yu-Lun Wu. Remote sensing image classification with the sen12ms dataset. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-2-2021, pp. 101–106, 2021. doi: 10.5194/isprs-annals-V-2-2021-101-2021.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, pp. 153–160, 2019. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE, 2019.

Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.