IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641144

A TRANSFORMER APPROACH FOR MULTI ORBIT PER PIXEL TIME SERIES FOREST CHARACTERIZATION WITH SENTINEL-1

Markus Zehner[1], *Valentin Kasburg*[2], *Clémence Dubois*[1,3], *Christian Thiel*[3], *Alexander Brenning*[4], *Jussi Baade*[5], *Nina Kukowski*[2], *Christiane Schmullius*[1]

[1]Friedrich Schiller University Jena, Department of Earth Observation, Jena, Germany
[2]Friedrich Schiller University Jena, Insitute of Geosciences, Jena, Germany
[3]German Aerospace Center (DLR), Institute of Data Science, Jena, Germany
[4]Friedrich Schiller University Jena, Department of Geoinformatics, Jena, Germany
[5]Friedrich Schiller University Jena, Department of Physical Geography, Jena, Germany

ABSTRACT

The Sentinel-1 (S-1) microwave measurements pose a unique opportunity for estimating forest parameters from satellite time series with increased data availability from overlapping orbits. Leveraging data from different orbits comes with varying viewing geometries and acquisition schedules, which can be considered in artificial neural networks.

We adapt a transformer architecture for mapping the full S-1 data against median forest height values. By doing so, we propose per-orbit temporal encoders to handle different acquisition times by position, missing data by attention masking, and the addition of viewing geometry context.

We show that our adjustments improve performance, with a greater impact of the additional data and a slight improvement in the masking. By optimizing the hyperparameters, our proposed method achieves a preliminary RMSE of 5.9 m and an rRMSE of 35 % in predicting the per-pixel vertical median forest height.

***Index Terms*—** SAR, SENTINEL-1, MULTI ORBIT, TRANSFORMER

1. INTRODUCTION

The temporal dynamics of S-1 have great value in providing timely interpolation on forest parameters such as top height, mean height, cover, distribution, and density derived from aerial laser scanning (ALS), as the satellite-based sensors excel in temporal-spatial coverage. However, considerations must be made based on the employed model because of different viewing angles and acquisition times. In recent studies on the estimation of forest parameters, S-1 is joined with optical imagery from Sentinel-2 (S-2) in a U-Net, concluding inferior performance of only S-1 and a minor increase in accuracy from S-2 only when adding S-1 [1]. While topographic artifacts and the speckle effect are named as challenging, the temporal frequency of S-1 time series is reduced when aligned with the less frequent S-2 acquisition dates. Similarly, S-1 showed poor variable importance as temporal statistics from leaf-on acquisitions in a Random Forest (RF) approach interpolating GEDI variables [2]. In contrast, [3] showed the value of the temporal dimension of S-1 data with a Long Short-Term Memory (LSTM) approach, which utilizes 97 scenes over five years to characterize forest metrics from S1 alone.

Recent works comparing Random Forest (RF), Convolutional- and Recurrent Neural Networks (CNN, RNN), and transformer architectures in mapping field crops with S-2 demonstrated the potential of using raw time series. For example, [4] achieved good performance and merits in preprocessed data, as well as the ability of an RNN and transformer model to detect and ignore missing information, such as cloud obstruction in optical images.

In this study, we strive to leverage the dynamics of all available overlapping orbits within a deep-learning approach to estimate the median forest height from ALS point clouds. In the context of computation cost, we adapt a Lightweight Transformer Attention Encoder (LTAE) architecture [5] to leverage the full S-1 time series and propose alterations to its specific challenges:

- splitting the variables so that each attention head is fed into one of the orbit's data,
- per-head time positions to take into account each orbits' acquisition scheme by hour-of-year,
- masking in the attention mechanism for missing values in the time series,
- and the addition of context to the sensor's perceived topography.

2. DATA AND METHODS

The dataset used for this study comprises Sentinel-1 (S-1) and ALS data from 2017 spanning the southern Hainich National

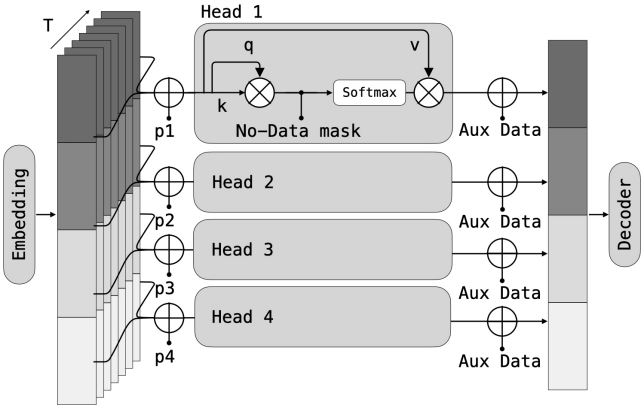

Fig. 1. Architecture of the LTAE adapted with per-head position (p1-p4), attention mask to handle missing data and adding auxiliary data before concatenation.

Park (HNP), Thuringia, Germany, which consists of mostly deciduous beech forest with few small patches of conifers. S-1 was processed locally to $\gamma_{0\,RTC}$ at 10 m spacing with a DEM from aerial laser scanning (ALS, ©GDI-Th, Freistaat Thueringen, TLBG) data as this showed better consistency of data from different orbits [6]. The S-1 data consists of simultaneously measured VH and VV polarized backscatter over 4 orbits with, on average, 60 acquisitions per orbit with an interval of 1 image/6 days, gathered by S-1A and B, with each orbit at shifted phases. The ALS data served as reference data, from which the per-pixel vertical median vegetation height with values ranging between 0 m and 40 m with a mean of 17.5 m were calculated. Sampling was done in a spatially separated fashion to have spatially distinct areas without overlap in the auxiliary data for training, validation and holdout samples. A constraint was sampling data with less than 30 missing values per pixel in VH or VV. After stratification and binning, the data was divided into training, validation, and test data sets with respectively 60%, 20%, and 20%. Due to the small sample size in the three highest stratification bins, the considered vertical median forest height ranges from 0 m to 34 m. The time series and vertical median forest height were scaled between 0 and 1, missing values were set to -1, and the time stamp of missing values was set to 0.

We chose the LTAE architecture because of its lower number of neurons within neural networks, which lessens the generally high computational cost compared to established machine learning methods. Along with efforts to reduce model size in the LTAE by using the mean query instead of calculating it for each time step [5], we saw the splitting of input variables on several low-dimensional heads to process each variable in a separate attention mechanism fitting for the data characteristics at hand.

We made several adjustments to the model proposed in [5] for the application of single-pixel regression of S-1, as depicted in Figure 1:

- embedding was changed to a continuous value embedding (CVE) [7] for the input of pixel time series,
- the position, which gives the transformer context of the temporal sequence of the input values, was adjusted to be separate in each head to enable processing of different orbits,
- a per-head attention mask was added into the scaled dot product to allow missing values in the input data,
- the pixel area was added as auxiliary data for the context of the viewing geometry for each sample by a 3x3 window embedded in a 2D convolutional layer.

Table 1. Parameters covered by the gridsearch.

Parameter	Values
Attn masking	True, False*
Auxil. data	True*, False
In Dim	4, 8, 16*
N Head	4, 8, 16*
H Dim	4, 8*, 16
Dropout	0.0*, 0.1
Loss Fn	RMSE, Huber loss*
Batch size	64*, 128, 256

The hyperparameter grid search was carried out with a fixed learning rate at 0.001 over values given in Table 1 using three differently sampled datasets. After 30 epochs, the models were evaluated via the RMSE on the validation data. The best-performing parameters were then used to train a final model with 100 epochs to evaluate against the holdout data.

3. RESULTS

The hyperparameter search resulted in 3888 models, whose performance in predicting median forest height was measured RMSE against the validation data in Figure 2 shows that the masking only marginally impacts performance, while the addition of the pixel area as auxiliary data shows visible improvement. Within the best 100 scores, 51 were achieved by applying the mask, whereas all utilized the auxiliary data. Increasing the dimensionality of the model by $In\,Dim$ of the CVE and the number of heads generally improves the performance. At the same time, the $H\,dim$ shows only a slight increase from 4 to higher numbers. Further, RMSE as a loss function, dropout of 0, and smaller batch sizes lead to better performance.

Using the best parameters marked with (*) in Table 1 achieved an RMSE of 5.9 m and an rRMSE of 35 %, predicting the vertical median tree height compared to a holdout dataset in Figure 3.

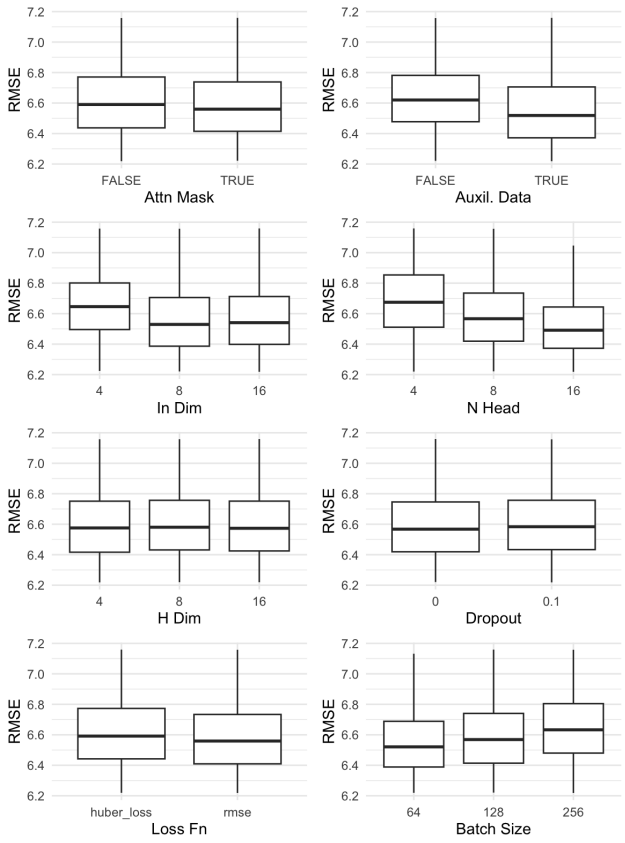

Fig. 2. Comparison of the grid search parameters, with the adaptations of attention masking and the factor of the auxiliary data in the first row, followed by dimensional parameters and general training parameters of the loss function and batch size. All compare the 3888 iterations split by the choice of the respective parameter. The Y-axis is limited between 0.1 and 0.9 percentile for better visibility.

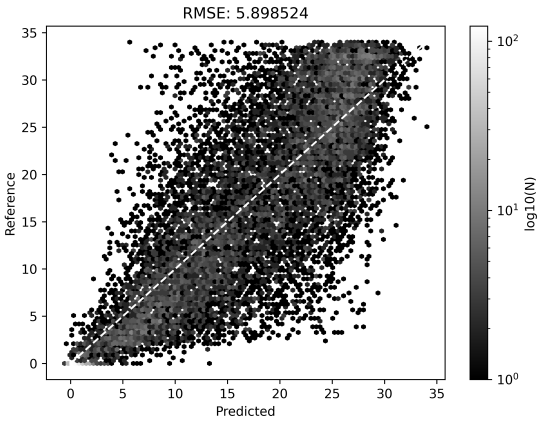

Fig. 3. Evaluation of per-pixel holdout-predictions versus reference data with the best parameters from the grid search. Lighter shades display more pixels falling into hexagonal bins in the plot coordinates.

The low impact of the mask can be attributed to the above-mentioned sample filtering to prefer samples with less than 30 missing values. Additionally, missing values are distinct without employing the attention mask, with values set to -1 and separated by the position argument set to 0. Including the auxiliary data led to an improved performance for the investigated parameters. However, besides providing additional information, the auxiliary factor also contributed to increased model size, which can be attributed to the embedding via the convolution layer. The dimension of the model is first set by $In\,Dim$, controlling the number of dimensions for each variable embedded by the CVE. The input dimension of the transformer follows then $In\,Dim * Orbits * Variables$ with $Orbits = 4$ and $Variables = 2$. Higher dimensions and a larger number of heads $N\,heads$ lead to a more flexible model and better performance on the training data but also increase the risk of over-fitting. When $N\,Head$ is set to 4, the model lacks the flexibility to generalize effectively on the data. Conversely, when $N\,Head$ increases from 8 to 16, the enhanced model flexibility yields marginal performance improvements. Dropout shows a slight negative impact as a measure against over-fitting that randomly blocks weights in the network. The lower performance of the Huber loss is probably due to the lower penalization of larger errors compared to using RMSE as a loss function.

The final model clusters values along the 1:1 line with a tendency to over-estimate vertical median forest height and finally under-estimate predictions above 25 m, seemingly saturating at high values. The results are prone to the displayed scattering with the current per-pixel approach.

4. DISCUSSION

We show the preliminary results of a transformer architecture to integrate overlapping orbits of S-1. Improvements to the current setup could be tackled by a longer time series as input instead of just one year, pre-training of the models [8]. Further, as shown in [4], preprocessing improves performance, which can also be included in the workflow, for example, a temporal smoothening by decomposition of the time signal [9]. Another approach for separating variables in time series data is taken in [7] by triplet embedding of time-variable-value while this complicates the later adding the per-orbit aux data. With respect to the dataset, spatial aggregation is expected to remedy the current straying in the evaluation, and reasonable trade-offs between spatial resolution and performance are to be explored.

While this work presents intermediate results within a spatially small dataset of the relative performance of our adjustments to the architecture, the proposed changes are not

limited to the transformer architecture. They are currently implemented in RNN and CNN to gain a broader comparison of the proposed adaptations. The current setup supplying acquisition time and missing data masks per variable and per sample also opens up the straightforward combination of raw time series from multiple satellite sensors, which suits the increasingly data-driven field of remote sensing.

5. REFERENCES

[1] A. Becker, S. Russo, S. Puliti, N. Lang, K. Schindler, and J. D. Wegner, "Country-wide retrieval of forest structure from optical and SAR satellite imagery with deep ensembles," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 269–286, 2023.

[2] P. Kacic, F. Thonfeld, U. Gessner, and C. Kuenzer, "Forest structure characterization in Germany: novel products and analysis based on GEDI, Sentinel-1 and Sentinel-2 data," *Remote Sensing*, vol. 15, no. 8, pp. 1969, 2023.

[3] S. Ge, W. Su, H. Gu, Y. Rauste, J. Praks, and O. Antropov, "Improved LSTM model for boreal forest height mapping using Sentinel-1 time series," *Remote Sensing*, vol. 14, no. 21, pp. 5560, 2022.

[4] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS journal of photogrammetry and remote sensing*, vol. 169, pp. 421–435, 2020.

[5] V. S. F. Garnot and L. Landrieu, "Lightweight temporal self-attention for classifying satellite images time series," in *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*. Springer, 2020, pp. 171–181.

[6] M. Zehner, C. Dubois, C. Thiel, K. Schellenberg, M. Rüetschi, A. Brenning, J. Baade, and C. Schmullius, "Accounting for deciduous forest structure and viewing-geometry effects improves Sentinel-1 time series image consistency," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[7] S. Tipirneni and C. K. Reddy, "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 6, pp. 1–17, 2022.

[8] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z. Zhou, "Sits-former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, pp. 102651, 2022.

[9] F. Cremer, M. Urbazaev, C. Berger, M. D. Mahecha, C. Schmullius, and C. Thiel, "An image transform based on temporal decomposition," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 537–541, 2018.