MULTI-BRANCH DEEP LEARNING MODEL FOR DETECTION OF SETTLEMENTS WITHOUT ELECTRICITY

Thomas Di Martino, Maxime Lenormand, Elise Colin Koeniguer

MULTI-BRANCH DEEP LEARNING MODEL FOR DETECTION OF SETTLEMENTS WITHOUT ELECTRICITY

Thomas Di Martino[1,2]*, Maxime Lenormand, Elise Colin Koeniguer*[2]

[1] SONDRA, ONERA, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
[2] ONERA, Traitement de l'Information et Systèmes, Université Paris-Saclay, 91123 Palaiseau, France

ABSTRACT

We introduce a multi-branch Deep Learning architecture that allows for the extraction of multi-scale features. Exploiting the data multi-modality structure through the combined use of various feature extractors provides high performance on data fusion tasks. Furthermore, the representation of the multi-temporality of the data using sensor-specific 3D convolutions with custom kernel size extracts temporal features at an early computation stage. Our methodology allows reaching performance up to 0.8876 F1 Score on the development phase dataset and around 0.8798 on the test phase dataset. Finally, we demonstrate the contribution of each sensor to the prediction task with the design of data-focused experiments.

Index Terms— Deep Learning, Multi Temporal, Remote Sensing, Multi Sensor, Classification, Data Fusion Contest

1. INTRODUCTION

In a Big Data context regarding the field of Earth Observation, a wide variety of satellite sensors become available. This diversity allows for the creation of multi-sensor application: thoroughly reviewed in [1], multi-modal classification of remote sensing images has shown its potential in agricultural [2] and urban [3] contexts.
In addition to the increased number of sensors, satellite image time series become available in a much higher temporal resolution, with shorter revisit times. In this context, a variety of studies [4, 5] demonstrated the usefulness of remote sensing time series to characterize and classify various environments temporally. Both optical [4] and radar [5] were shown to contain critical temporal information for applications such as land cover classification or crop monitoring.
The convergence of multi-temporality and multi-modality in remote sensing applications is a matter of interest and has been studied in [6] where the fusion of High Spatial Resolution and Very High Spatial Resolution satellite image time series was computed to classify land cover.

2. DATASET PRESENTATION

Fitting into this mix of multi-modality and multi-temporality, the dataset of the Data Fusion Contest for detection of settlements without electricity (DFC21-DSE) [7], developed in parallel of Solaraid's objectives to electrify isolated regions of Africa, consists of images from:

- Sentinel 1-A GRD (S1), a 5x20m resolution radar sensor (VV and VH polarisations) resampled to 10x10m resolution, for four dates;
- Sentinel 2 (S2), 12 channels of reflectance with ground sampling distances of 10m, 20m, and 60m, for four dates;
- Landsat 8 (LC08), 11 channels with ground sampling distances of 15m, 30m and 100m, for three dates;
- VIIRS, with only Day-Night Band, originating from the VNP46A1 dataset that provides nocturnal visible and NIR light measurements at a GSD of 750m resampled at 500m, for nine dates;

With a total of 98 images split as 60 training, 19 validation, and 19 test images, all resampled to 10m resolution, the dataset is supplied with labels of a resolution of 500m, as 16-by-16 images, including four classes, among which only one has to be retrieved for algorithm evaluation: Human settlements without electricity. The distribution of the four classes is shown in table 1.

Table 1: Classes distribution with color codes

	With **electricity**	Without **electricity**
With **settlements**	676	(ROI) 6318
Without **settlements**	211	8155

Designed initially as a semantic segmentation task, as illustrated in Fig. 1, we decided to take advantage of the spatial

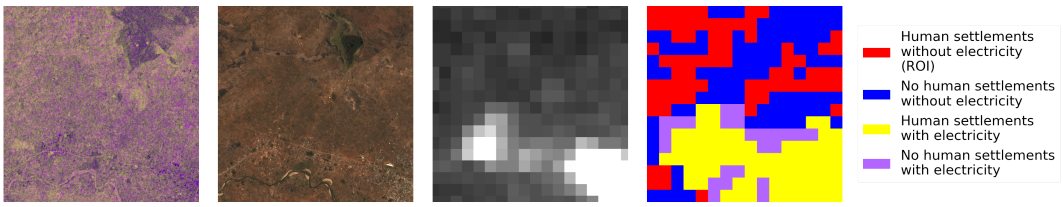

Fig. 1: Dataset Extract (Training Tile #14). Left to right: Sentinel-1 (VV, VH, VV-VH), Sentinel-2 (B04, B03, B02), VIIRS averaged, Labels

resolution of the image labels: a single class maps areas of 50-by-50 pixels. We assume that a sufficiently negligible amount of information lies in the spatial arrangement of classes within the 800-by-800 class arrangement so that splitting the image into tiles does not impact predictions. Thus, we transformed the problem into a classification problem where, given an input image of size (50,50,98), we want to predict its class among the four available categories. Hence, we split each 800-by-800 image into 256 50-by-50 tiles. We obtain $15,360$ labeled samples, $4,864$ validation and $4,864$ test samples.

3. METHODOLOGY

3.1. Multi-Branch Multi-Temporal Architecture

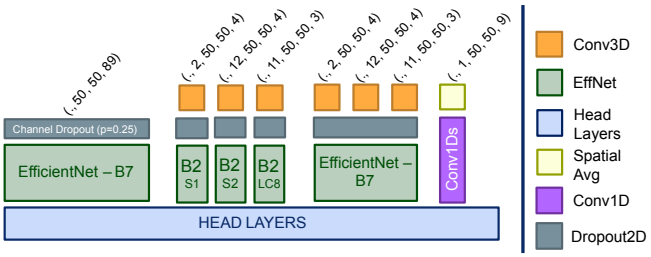

Fig. 2: Introduced Multi-Branch Multi-Temporal architecture

Our model architecture is a Multi-Branch network, as seen in Fig. 2. Its design incorporates the multi-temporal and multi-modal aspects of the data.

3.1.1. Multi-temporal Feature Extraction

To tackle the dataset's multi-temporality, we use a Conv3D layer as the first step in most of the branches' computation. It is designed to reduce the temporal component early on in the model computation: we do not consider the temporal dimension of the data as crucial for the prediction but rather as a way to limit data noise.
Thus, given a data stack of shape (c, h, w, t) where c corresponds to the number of channels of the stack, h its height, w its width and t its temporal dimension, we design a Conv3D layer with a kernel of size $(t, 1, 1)$, and c filters so that it reduces only the temporal dimension. The reduced stack is then passed on to Conv2D layers for spatial feature extraction.

3.1.2. Multi-modal Feature Extraction

The architecture of our model consists of four feature extraction branches:

- **DNB-specific TempCNN Branch**: given the low resolution of the VIIRS sensor, equal to the label resolution, our model only extracts temporal features using a stack of 1D Convolutions, after we spatially averaged the band's value over the 50-by-50 image area.

- **Multi-modal Branch**: the leftmost architecture, presented in Fig. 2, takes as input a batch of images and does not encode any prior information regarding the structure of the input data (for example, difference in sensors). This branch aims at extracting features that span across any acquisition and any sensor.

- **Multi-Unimodal Branch**: the center model is also a multi-branch model that considers both the multi-sensor and multi-temporal aspects of the input data. For that matter, we have trained three conceptually identical sub-branches, each with data specific to a single sensor. This *blindfolding* strategy aims at retrieving as much information as possible from each sensor separately. These sensor-specific branches are built using the same components: we first have a Conv3D layer as presented in Section 3.1.1. The extracted temporal features are then passed to a Dropout2D layer before an EfficientNet B2 [8].

- **Temporal-Merged Branch**: This *Temporal-Merged* branch performs the same temporal reduction as the Multi-Unimodal model. The temporally-reduced stack of each sensor data are then concatenated, channel-wise, and fed to an EfficientNet B7 [8].
  This design removes the potential temporal redundancy within each sensor data before extracting features that may span across multiple sensors.

After features extraction from each of the model's branches, we flatten and concatenate them to a common feature vector which is then passed to a stack of fully connected layers.

3.2. Environment

As presented in Fig. 3, our model fits into a custom training environment. We empirically designed it to match the requirements of the task and of the dataset.

Loss: We combine in our loss function a Categorical Cross-Entropy (CCE) function to extract semantic information specific to each of the four classes as well as a Soft-F1 Loss function to focus the learning on the correct classification of the class of interest, in a 1-vs-all fashion.

Augmentation: As the amount of data is relatively small, we decided to opt for a heavy augmentation strategy. We use Flips (Horizontal + Vertical), Rotations (mod $\pi/2$), Noisy-Labels with randomly shifted cropping, same-class cut mix [9], with an augmentation factor of 16.

Ensembling: The training set is randomly split into three Folds, implying three separately trained models, one for each fold. Every model are then ensembled and we perform test-time augmentation before averaging their prediction.

4. EXPERIMENTAL DESIGN

The biggest challenge of the competition is to be able to leverage each of the available modalities. To measure their contribution to the final prediction, we designed two sets of experiments. In the context of the DFC21-DSE dataset [7], we can consider the four classes as the intersection of two binary classes, as presented in table 1: first, we need to detect if any settlement is present in the 50×50 image. Then, we need to analyze whether the region is electrified. Considering the sensor data provided, we assume that optical and radar sensors tend to focus on the detection of settlements and that the VIIRS sensor assesses the electrification status of the region. The fusion of SAR and optical features were shown to have high performance and complementarity when applied to building detection [10]. Regarding VIIRS sensor data, preliminary results of the application of VIIRS DNB data to detect a power outage in India [11] show encouraging results for the application of VIIRS sensor data to the task.

4.1. VIIRS Day/Night Band for electricity detection

When assessing VIIRS DNBs' ability to detect electrified regions on our dataset, we obtain the results displayed in table 2. To generate these results, we isolate the DNB-specific TempCNN branch of our model to classify if a 50-by-50 image is electrified (positive class) or not (negative class). We can conclude that a correlation exists between VIIRS DNB data and the presence of electricity within a scene. Hence, this sensor is crucial to classify non-electrified settlements.

Table 2: VIIRS training for electricity detection F1 score

Subset	Fold 1		Fold 2		Fold 3	
	Train	Val	Train	Val	Train	Val
VIIRS	0.712	0.718	0.765	0.462	0.674	0.75

4.2. Assessment of optical and radar contribution to the final prediction task

To evaluate the contribution of the other 3 sensors at this task, we studied their mutual contribution to the final prediction by designing series of data-focused experiments where we select a subset of the available sensor data to comparatively evaluate their respective performance. Two conclusions can be made from the results displayed in table 3. The first, least surprising, is that the use of all data available provides the highest performance of all sensor combinations. The second is the apparent superiority of S2 which can be linked to multiple factors: S2's higher resolution than LC08, especially in visible light bands, provides it with better tools to detect isolated and small settlements; the dimension of buildings of around 10m makes them difficult to perceive within the speckle of S1; However, the addition of SAR interferometric information, with, for instance, S1 Single-Look Complex imagery, would probably improve its contribution to the classification task. Hence, the task of assessing electrification of isolated regions may benefit both from advanced Deep Learning models and the use of more expert remote sensing data.

Table 3: Data Subset experiments with local validation F1 score

Subset	Fold 1		Fold 2		Fold 3	
	Train	Val	Train	Val	Train	Val
S1, VIIRS	0.681	0.653	0.657	0.665	0.677	0.672
LC8, VIIRS	0.766	0.758	0.750	0.745	0.768	0.776
S2, VIIRS	0.834	0.817	0.822	0.824	0.850	0.859
S1, S2, LC8, VIIRS	0.893	0.854	0.900	0.853	0.878	0.872

5. FINAL RESULTS OF THE MULTI-BRANCH ARCHITECTURE

Table 4: Winning model submission F1 Score

Fold 1 Val	Fold 2 Val	Fold 3 Val	Dev Phase	Test Phase
0.8547	0.8533	0.8722	0.8877 (1st)	0.8798 (3rd)

When combining the training strategy presented in Fig. 3, the multi-branch model used in Fig. 2 and every sensor data available, we obtain the results presented in table 4.

6. CONCLUSION

In this paper, we present our winning method for the DFC21-DSE 2021 by detailing the custom model built for the task and the training and inference setup used for the model. Besides,

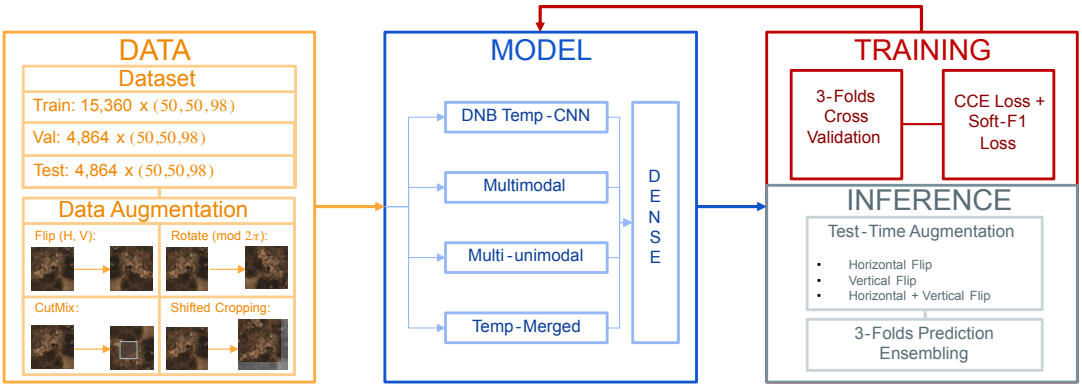

Fig. 3: Description of the model training and inference environment

we also explore each sensor contribution to detecting settlements without electricity, with the objective that our results may generalize to other regions of Earth. With the presented method, we obtain an F1 Score of 0.8877 on the development phase dataset and 0.8798 on the test phase dataset, thus ranking us at the 3rd place of this competition.

6.1. Acknowledgement

The authors would like to thank the IEEE GRSS Image Analysis and Data Fusion Technical Committee, Hewlett Packard Enterprise, SolarAid, and Data Science Experts for organizing the Data Fusion Contest. We also thank ONERA's IVA team for their support during the challenge, especially Adrien Chan Hon Tong, Aurélien Plyer and Guy Le Besnerais.

7. REFERENCES

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[2] A. Lapini, G. Fontanelli, S. Pettinato, E. Santi, S. Paloscia, D. Tapete, and F. Cigna, "Application of deep learning to optical and SAR images for the classification of agricultural areas in italy," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 4163–4166.

[3] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 6, pp. 1709–1724, 2019.

[4] Charlotte Pelletier, Geoffrey I. Webb, and François Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, 2019.

[5] M. Lavreniuk, N. Kussul, and A. Novikov, "Deep learning crop classification approach based on sparse coding of time series of satellite data," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 4812–4815.

[6] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "m^3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, pp. 4939–4949, 2018.

[7] Naoto Yokoya, Pedram Ghamisi, Ronny Hansch, Colin Prieur, Hana Malha, Jocelyn Chanussot, Caleb Robinson, Kolya Malkin, and Nebojsa Jojic, "2021 data fusion contest: Geospatial artificial intelligence for social good [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 287–C3, 2021.

[8] Mingxing Tan and Quoc V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[9] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," 2019.

[10] H. Sportouche, F. Tupin, and L. Denise, "Building detection by fusion of optical and SAR features in metric resolution data," in *2009 IEEE International Geoscience and Remote Sensing Symposium*, 2009, vol. 4, pp. IV–769–IV–772.

[11] Michael L. Mann, Eli K. Melaas, and Arun Malik, "Using VIIRS day/night band to measure electricity supply reliability: Preliminary results from Maharashtra, India," *Remote Sensing*, vol. 8, no. 9, 2016.