Explaining the decisions and the functioning of a convolutional spatiotemporal land cover classifier with channel attention and redescription mining

Enzo Pelous, Nicolas Méger, Alexandre Benoit, Abdourrahmane Atto, Dino Ienco, Hermann Courteille, Christophe Lin-Kwong-Chon

Explaining the Decisions and the Functioning of a Convolutional Spatiotemporal Land Cover Classifier with Channel Attention and Redescription Mining[⋆,⋆⋆]

Enzo Pelous[a], Nicolas Méger[a,*], Alexandre Benoit[a], Abdourrahmane Atto[a], Dino Ienco[b,d], Hermann Courteille[c], Christophe Lin-Kwong-Chon[a]

[a] *Université Savoie Mont Blanc, Polytech Annecy-Chambery, LISTIC, 5 chemin de Bellevue, Annecy-le-Vieux, 74940, Annecy, France*
[b] *INRAE, UMR TETIS, Université de Montpellier, 500 rue Jean-François Breton, Montpellier, 34000, France*
[c] *Univ. Rennes, INSA, CNRS, IRISA, Campus de Beaulieu, Rennes, 35042, France*
[d] *INRIA, Université de Montpellier, 161, rue Ada, Montpellier, 34000, France*

Abstract

Convolutional neural networks trained with satellite image time series have demonstrated their potential in land cover classification in recent years. Nevertheless, the rationale leading to their decisions remains obscure by nature. Methods for providing relevant and simplified explanations of their decisions as well as methods for understanding their inner functioning have thus emerged. However, both kinds of methods generally work separately and no explicit connection between their findings is made available. This paper presents an innovative method for refining the explanations provided by channel-based attention mechanisms. It consists in identifying correspondence rules between neuronal activation levels and the presence of spatiotemporal patterns in the input data for each channel and target class. These rules provide both class-level and instance-level explanations, as well as an explicit understanding of the network operations. They are extracted using a state-of-the-art redescription mining algorithm. Experiments on the Reunion Island Sentinel-2 dataset show that both correct and incorrect decisions can be explained using convenient spatiotemporal visualisations.

Keywords: Explainable AI, Convolutional Neural Networks, Land Cover Classification, Satellite Image Time Series, Attention, Redescription Mining, Grouped Frequent Sequential Patterns.

1. Introduction

Satellite Image Time Series (SITS) are spatiotemporal data acquired by satellite missions such as the Landsat and Sentinel ones, which even provide free access to their archives. This context favours the use of SITS for change detection and Land Cover Classification (LCC), with applications ranging from disaster management to agricultural monitoring or urban growth assessment. These tasks have also benefited from the rise of deep learning techniques in recent years. In particular, Convolutional Neural Networks (CNNs) have been shown to reach high classification performances (e.g., [1]).

Like any deep learning architecture, CNNs consist of thousands, if not millions, of parameters. As a result, the reasoning behind their decisions is often opaque. *Explainable Artificial Intelligence (XAI)* [2] has therefore emerged with the aim of providing prediction explanations that are simple enough to be understood by end users. They generally point out the main data components taken into account in generating the decisions. One can distinguish post-hoc methods such as GradCam [3] that provide prediction explanations of complex black box models and intrinsic methods that aims to create understandable models by design [4]. The latter can be achieved, for example, by using attention operators [5]. In addition to providing low-level explanations of the predictions, these operators can also improve task performance through regularization. The precise description of the way in which the main data components are processed within the networks is beyond the scope of the explanatory methods and is left to *interpretation methods*. [6]. The latter can, for example, identify the parts of the network that perform wavelet filtering and characterize the wavelet bases that are learned.

However, these two types of methods typically operate separately, with no direct link between their results. Going further, this paper proposes to combine both types of methods in order to improve model understanding, and not model performance which is left to other contributions. Revising the dataset or the network according to interpretations and explanations in order to improve its

⋆This work was supported by CNES, focused on the MultiSpectral Instrument and on the Sentinel-2 mission.

⋆⋆This work is supported by the ANR REPED-SARIX project (ANR-21-CE23-0012-01) of the French national Agency of research.

*Corresponding author

Email addresses: enzo.pelous@univ-smb.fr (Enzo Pelous), nicolas.meger@univ-smb.fr (Nicolas Méger), alexandre.benoit@univ-smb.fr (Alexandre Benoit), abdourrahmane.atto@univ-smb.fr (Abdourrahmane Atto), dino.ineco@inrae.fr (Dino Ienco), hermann.courteille@irisa.fr (Hermann Courteille), christophe.lin-kwong-chon@univ-smb.fr (Christophe Lin-Kwong-Chon)

performance is also not addressed in this paper. Nevertheless, some methodological directions are identified and presented. The proposed method was originally sketched in [7]. It consists in refining the channel attention-based explanations of a CNN model with interpretations that are explicitly associated with explanations. These associations are expressed as correspondence rules between neuronal activation levels and the presence of spatiotemporal patterns in input data. They are extracted for each channel and each class using a state-of-the-art redescription mining method, namely the *ReRemi* algorithm [8]. Validation experiments address an LCC task using a Sentinel-2 SITS acquired over the *Réunion* island. In such a task, experienced end users typically use their knowledge of target classes and input data channels. Following a similar approach, our proposal allows explaining predictions by providing global information about the channels most involved in the decision by using an attention module, while extracted correspondence rules give access to the details of the detected activation patterns. The contributions of this paper thus concern a channel attention-based CNN performing LCC and are to 1) refine attention-based explanations by providing complementary interpretations and explanations using a single concept, namely redescriptions, 2) propose human-readable interpretations and explanations under the form of neuronal activation levels and spatiotemporal pixel evolutions, and 3) demonstrate the interest of the approach on a real SITS.

This paper is structured as follows: CNN-based LCCs learnt from SITS as well as explanation and interpretation methods are reviewed in Section 2. A background on pattern mining is then made available in Section 3: (1) the spatiotemporal patterns employed to describe the input data in an unsupervised way are presented in Section 3.1, and (2) the concept of redescriptions used to extract correspondence rules between these patterns and neuron activation levels is detailed in Section 3.2. Section 4 introduces the proposed approach in terms of general workflow and network architecture. After having presented the Sentinel-2 dataset exploited in this paper in Section 5.1 and the experimental settings in Section 5.2, quantitative and qualitative results are respectively provided in Section 5.3 and Section 5.4. Section 6 discusses the proposed method in terms of design and usage. Finally, Section 7 concludes this paper by outlining the potential and the limitations of the proposed method that ground our future work directions.

2. State of the art

2.1. Land cover classification with convolutional neural networks learnt from SITS

The rise of deep learning techniques for land cover classification is driven by their performances but also by their ability to automatically learn in an end-to-end fashion the most suited data features [9]. They thus potentially avoid relying on hand-crafted descriptors such the Normalized Difference Vegetation Index (NDVI) proposed in [10] as that can lack of generality or specificity. In addition, both linearity and non-linearity can be learnt [11]. Also, optimizing a classifier from SITS is a challenging task since SITS present extremely rich information expressed along the spatial, temporal and spectral dimensions. Still, various architecture types such as Recurrent Neural Networks (RNNs) (e.g., [12] or [13]) or Convolutional Neural Networks (CNNs) (e.g., [1]) have been proven to perform well on such data. In this paper, as explainability and interpretability are seeken, CNNs are focused on. Convolutions are indeed interpreted directly as filters, while the field of view (FOV), i.e. the extent of influence of the input neighbourhood, can be controlled by design. In this paper, the FOV is set to match the spatio-temporal extent of the patterns used for explanations. It takes into account both the number of acquisitions and the spatial autocorrelation.
Regarding the neural network model architecture, convolutions performed at the pixel level along the temporal axis have been shown to produce good results such as those obtained in [14], [1] or [15]. The difference between these works lies in the fact that a single vegetation index built using different channels is used in [14], while convolutions are applied to all available channels in [1] and [15] without any a priori. Simple 2D convolutions performed along the spatial dimension is far from being sufficient, the temporal dimension being of primary interest as evidenced in [14] or [1]. Approaches incorporating, in a same network, both a temporal analysis performed through RNN layers and a spatial processing achieved with 2D convolutional layers have thus been proposed in [16] or [17]. In that case, each channel is separately processed before aggregating features and predicting classes. Another approach proposed in [18] or [19] consists in feeding each recurrent cell of a RNN with the spectrospatial data observed at the different acquisition dates. In [19], spectrospatial patches are simply flattened as 1D vectors that are then supplied to standard recurrent cells while the latter are modified in [18] to perform spectrospatial convolutions on input data directly. The spatial dimension can also be handled at the object/region level by extracting them through clustering, establishing a representative sequence of each object for each channel, and finally convolutioning all of the sequences temporally before the final classification stage [20]. Finally, if the time and space dimensions are regularly sampled, performing 3D convolutions is appropriate if one expects to identify relevant local spatiotemporal patterns in each data channel separately. For instance, this approach was adopted to build up feature extractors in the first model layers before the channel features fusion step and classification heads in [21] or [7].
In general, most of the previously cited work remains black box models with no or limited explanatory or interpretative behaviour. Some models, such as [7], involve attentional processes to provide model regularization and prediction explanation with respect to input data channels,

but performance improvement is generally preferred over prediction justification. As will be shown in the next sections, the latter direction can nevertheless be explored, taking advantage of recent advances in the explanability and interpretability of deep neural networks.

2.2. Explainability

Along the continuous progress of deep neural networks, their integration into safety-critical applications is gaining attention. However challenging case studies such as autonomous driving, robotics and medical diagnosis usually involve complex models that regulatory authorities now gradually impose to be explainable. In recent years, many research directions have been developed on this topic as described in [2]. As a brief summary, two main families of approaches can be distinguished and are applicable to CNN models as well as other model structures. First, the "post-hoc" methods allow for the extraction of explanations from already trained models. Such methods thus require additional processing after each model prediction. Among the most commonly used methods, one can cite the GradCam approach [3], dedicated to CNN models, that produces a heatmap in the input space showing the most contributing areas that led to a given prediction. The Layer-Wise Relevance Propagation (LRP) method [22] goes further by propagating prediction backward relying on refined rules and justified as a deep Taylor decomposition. This results in very detailed explanation maps that highlight the contribution of each input pixel. However, such family of methods generally requires access to internal operations within the models and often only report positive (excitatory) contributions to the prediction. Other model agnostic methods such as Local Interpretable Model-Agnostic Explanations (LIME) [23] and SHapley Additive exPlanations (SHAP) [24] can provide more details and highlight excitatory or inhibitory patterns. In [24], it is shown that SHAP improves over LIME by relying on Shapley values that enable local accuracy, missingness and consistency. However, such method makes the hypothesis of features independence and requires significant computations, which may limit its application.

Another explainability approach consists in directly integrating intrinsic constraints from the design of the model. This is also referred to as *self-explaining* or *interpretable-by-design* models [4]. One advantage is that explanation extraction does not need additional processing and is provided simultaneously with the prediction. Different directions are possible. If the problem is well identified, one can impose the model to detect specific attributes that make up the target concepts *prototypes* such that prediction can be explained with respect to some expected references. Usually models produce intermediate latent representations to be matched with prototypes. ProtoPNet [25] and ProtoTree [26] are typical examples of image recognition tasks. Such methods are relevant but may highlight some bias such as the "Clever Hans" [27] phenomenon that relates to favouring patterns highly correlated to the target but not expected in the design step. Typically, background patterns may systematically appear with a specific foreground target class and thus generate such bias. There is also a debate on the use of attention processes to build up self-explaining models [28]. At a given stage along a deep neural network, attention operators learn to automatically focus on some specific inputs to improve the target task. Intuitively, it identifies a regularity in the activation patterns associated to a given prediction and tries to mask other potential disturbing activation. Attention operators then highlight the contributing features associated to a prediction thus providing a rough explanation. Then, at the model design step, attention processes can be carefully placed within the model in order to provide end-users meaningful explanations. However, similarly to GradCam, provided explanation only report the fact that a given feature is involved but does not report on its influence (excitatory or inhibitory) nor its strength with respect to the final decision. Finally, introducing constraints related to the underlying physical models can also contribute to model explanation. Typically, the use of additional losses that qualify the relevance of a prediction or an intermediate feature with respect to a physical dimension has been reviewed in [29].

2.3. Interpretability

In interpretability, knowledge associated with upstream to downstream neuronal information exchanges are seeken. They are related to the transmission/blocking properties or the penalization form that can apply to any specific information carried by the input data or by some given features. The interpretability knowledge can be deduced by analysing the statistical properties learned by convolutional filters. In the same layer, the statistical characterization generally involves either studying the correlations between the convolutional filters (high correlations results in general in redundant information computation), or analysing the type of information on which the convolution kernels focus. A distinction is thus made in [30] between kernels which will operate weighted moving averages (meanlets, low-pass filters), kernels which will rather operate weighted moving differencials (differencelets, high pass filters) and other kernels which will distort (distortlets) different parts of the information contained in the data. In addition to this intra-layer based inter+intra-kernel characterizations, an inter-layer consideration can also be used to evaluate the decrease of randomness or the increase of entropy from upstream to downstream layers.

Alternatively, Physically Inspired Neural Networks (PINNs) introduce inner model design constraints to comply with knowledge on the considered physical phenomenon, which thus provide interpretability by design. Often considered in application domains relying on partial differential equations (PDEs) such as fluid dynamics, these methods, however, suffer from learning bias both related to the partial knowledge of the physical problem and spectral

bias that relate to the difficulty of models to learn high frequency functions [31].

3. Background on pattern mining

The proposed approach consists in refining the explanations conveyed by channel attention weights with supplementary interpretations and explanations that are provided by pattern mining techniques. The latter are envisaged to make as few assumptions as possible. They are indeed designed to support *Knowledge Discovery In Databases (KDD)* processes, which requires providing data descriptions that are not biased towards end users' expectations [32]. The proposed pattern-based interpretations and explanations are expressed as correspondence rules between neuronal activation and spatiotemporal patterns present in input data. These patterns, namely *GFS-patterns*, as well as the correspondence rules, namely *redescriptions*, are presented in Section 3.1 and Section 3.2 respectively.

3.1. GFS-patterns

The *Grouped Frequent Sequential patterns (GFS-patterns)* originally proposed in [33] are mined in a fully unsupervised way to identify which spatiotemporal regularities are present in input data. This is performed for each data channel separately by a) symbolizing original radiometric values, b) mining pixel-based symbolic evolution patterns and c) selecting the most interesting ones. Each one of these steps is described hereafter.

Symbolization. Symbols are obtained by quantizing radiometric values. For instance, utilizing an equal frequency bucketing based on the 33^{rd} and the 66^{th} percentiles, original pixel values can be quantized over three intervals. As a result, for each channel, pixel values are denoted with symbols '1', '2', and '3' to respectively represent low, medium and high radiometric values. The original SITS is thus transformed into a symbolic one.

Mining. Pixel evolution patterns and sub-patterns expressed as symbolic sequences such as $2 \rightarrow 3 \rightarrow 1$ are then mined. If the latter occurs in a symbolic pixel evolution sequence, then it indicates that, some time in the sequence, the symbol of the pixel it describes is '2', then, sometime later '3', and, finally, sometime later '1'. No timing constraint is imposed. A pattern is retained if 1) it affects a sufficient number of pixels, i.e., it covers a minimum surface denoted σ, and 2) these pixels are sufficiently connected to each other in their immediate 3×3 neighbourhoods, i.e., they form homogeneous regions whatever their shapes. Pattern occurrences are thus *frequent* and *grouped*, hence the name of *Grouped Frequent Sequential Patterns* or *GFS-patterns*. The reader is referred to [33] for a more formal definition of GFS-patterns and details regarding the corresponding extraction algorithm.

Selection. Only *maximal* GFS-patterns are filtered out to focus on the most specific ones, i.e. those that are not contained in any other pattern of the output collection. Finally, the maximal GFS-patterns that are the less or the more likely to occur in randomized versions of the symbolic datasets, i.e., the most interesting GFS-patterns, are retained. More details about this ranking method can be found in [34]. In the following, the most interesting maximal GFS-patterns are simply referred to as *patterns* when clear from the context.

3.2. Redescriptions

The pattern-based explanations and interpretations proposed in this paper relate neural activation levels observed at inference time to the presence of patterns in the input dataset. This is achieved for each class and each channel by describing classified pixels with the different activation levels of the neurons and GFS-patterns, and by unveiling correspondence rules between these two different types of descriptions. Let a_i denote the activation level of neuron i, $\wedge$ the logical *AND* and $\sim$ the correspondence between the left-hand side description and the right-hand side one. Expressions such as $0.2 < a_1 < 0.3 \wedge 0.7 < a_{17} < 0.9 \sim 2 \rightarrow 3 \rightarrow 1$ are thus targeted. The latter means that the pixels for which the activation levels of neurons a_1 and a_{17} respectively belong to $]0.2; 0.3[$ and $]0.7; 0.9[$ at inference time tend to be affected by pattern $2 \rightarrow 3 \rightarrow 1$, and vice versa. Such an expression is termed *redescription* by the data mining community. More generally, *redescription mining* is *'a data analysis task that aims at finding distinct common characterizations of the same objects'* [35].

Back to the explanation example, it is built upon two distinct classified pixel descriptions, $p = 0.2 < a_1 < 0.3 \wedge 0.7 < a_{17} < 0.9$ and $q = 2 \rightarrow 3 \rightarrow 1$. Description p originates from the table denoting the neural activation levels while q is produced according to the table reporting the presence of patterns. The set of objects for which a description is valid is termed *support*. A *redescription* is a pair of descriptions such as (p, q), also denoted $p \sim q$, each description being produced from a different table. In order to evaluate the accuracy of a redescription $p \sim q$, its Jaccard index is computed as $\frac{|supp(p) \cap supp(q)|}{|supp(p) \cup supp(q)|}$.

In this paper, the *ReReMi* algorithm proposed in [8] is considered. It can indeed automatically determine the optimal numerical intervals that are considered when establishing a description from numerical values such as the neural activation levels. These intervals are built on the fly to favour redescription accuracy. They can thus differ from one redescription to another. Among all possible redescriptions, algorithm *ReReMi* retains those whose Jaccard index exceeds a user-defined threshold and whose descriptions are statistically dependent. This dependence is checked using a p-value expressing the probability that the supports of descriptions overlap as much as observed. Such a significance test tends to favour redescriptions with low support and can be counterbalanced by rejecting those whose support is below a user-defined threshold.

4. The *PM4X* method

4.1. General workflow

The *Pattern Mining 4 eXplanations (PM4X)* method proposed in this paper is aimed at refining attention weights explanations with pattern-based ones. It chains three different steps that are depicted in Figure 1. The first two, steps 1a and 1b, are respectively dedicated to the learning of a pixel-based LCC to be explained and the extraction of GFS-patterns. Both operations are independent and can be fully parallelized. More specifically, it is assumed that the chosen LCC captures features that are meant to be similar to GFS-patterns. The latter are extracted seperately for each channel, considering all training partitions to induce patterns that are as general as possible. It is recalled that this extraction is fully unsupervised and performed after transforming the input SITS into a symbolic one (see section 3.1).

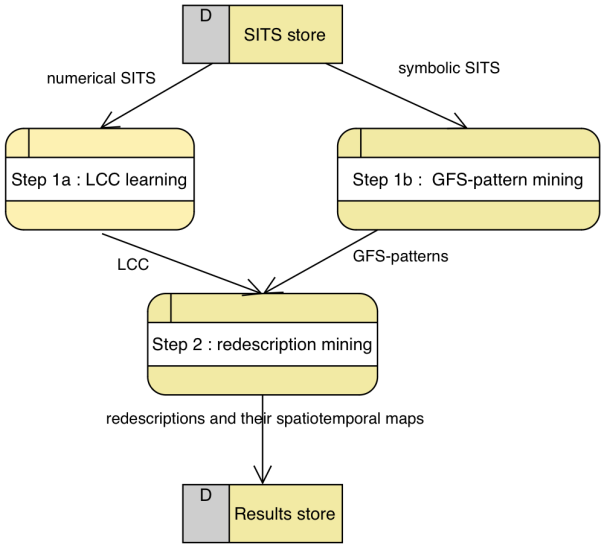

Figure 1: The *PM4X* data flow: input/output data/models are denoted by arrows, rounded boxes represent processes.

In Step 2, the pixels of each training partition are described with the neuron activation levels observed when inference is performed on original values, only for the neurons that are assumed to express features resembling GFS-patterns. Such features are obtained after having cascaded spatiotemporal and temporal convolutions. The exact layer gathering these features is described hereafter in Section 4.2. The same pixels are also described by checking whether or not patterns affect them in the symbolic version of the SITS. These descriptions are then mined to extract redescriptions, i.e. correspondance rules between neuron actication levels and patterns (see Section 3.2). Since all training partitions are considered, extracted redescriptions are as general as possible. This extraction is performed for each channel and each class, so extracted redescriptions can be considered as class-level explanations. They can also be used at the instance level, i.e. the pixel level, to explain each one of the decisions by checking whether they hold or not. Redescriptions can be visualized by relying on a variant of the *SpatioTemporal Localizations Maps (STL-maps)* [34] that were originally designed to visualize patterns both in space and in time. As shown in Section 5, these maps are convenient when it comes to understanding both right and wrong decisions. Finally, redescriptions can also be considered as interpretations of the inner functioning of the LCC, since they identify both the neurons mobilized when capturing input data patterns and their activation levels.

4.2. Network architecture

The eXplainable Deep Spatiotemporal Land Cover Classifier (*X-DSLCC*) architecture used in this paper is mainly inspired by those proposed in [1, 15, 7]. It is designed so as 1) to be interpretable as possible, and 2) compute features that match GFS-patterns as much as possible. Model performance is left to other contributions. As explained in Section 2.1, CNNs are to be favoured since their spatiotemporal FOV can be controlled by design and their convolutions can be simply interpreted as filtering. A CNN is thus considered. Its architecture is presented in Table 1. In more detail, layer ① separates each channel allowing for specific neural paths to specialize feature extraction for each of them. Each channel is then filtered in a spatiotemporal manner using layer ② to apply $N_1 = 256$ different convolution matrices at the pixel level. Their dimension is $k_1 \times 3 \times 3$ without applying padding. This FOV is chosen so as to match the one used to extract GFS-patterns. More precisely, the spatial footprint 3×3 corresponds to the spatial extent considered to check whether GFS-pattern occurrences are grouped or not. Once applied, since no padding is used, the 3x3 spatial input features are reduced to 1D signals along the temporal axis thus reducing local features and simplifying the next processing steps. Regarding the temporal FOV k_1, it is experimentally set to a third of the series length T to avoid compressing the temporal information in a too harsh manner. Pattern lengths can indeed range between 1 and T symbols. The latter precaution also holds regarding the following temporal convolutions. Instead of relying on the proposal available in [7] that chains two temporal convolution layers after the spatiotemporal one, a single convolution layer composed of $N_2 = 64$ temporal filters of size $k_2 = k_1 = T/3$ is considered with layer ③. Interestingly, for the datasets used in this paper, no performance degradation is to be reported when doing so. Model relevance is thus maintained while being more frugal. Obtained features are 1D vectors containing T_2 temporal values. They are assumed to match GFS-patterns and are stacked using layer ④. The neuron activation levels to be associated with patterns are therefore those observed at inference time at the output of layer ④. These features are then weighted thanks to a channel attention module referred to as layer ⑤ before reaching the decision layers that stacks ⑥, a hidden dense layer that

is ReLu activated and connected to the final dense one ⑦ that relies on the softmax function. Back to the attention operator, it allows dynamic weighting of the importance of each channel in the final decision of the network. It thus provides significant explanations, i.e., contribution of each input channel to a given decision. These weights can be easily merged for each class using box plots [15]. To this aim, we consider a simple attention process proposed in [36] formalized as $\alpha_i = sigmoid\left(< \vec{u}, tanh(W\vec{h_i} + \vec{b}) >\right)$ where α_i is the attention weight associated to channel i obtained from a non-linear transformation with learnt weights W (2D), u (1D) and bias b (1D) applied to the vectorized (flatten) features outing from layer ④.

To help the training process convergence, inputs are normalized channel-wise between 0 and 1 using a Min-Max scaling. The usual unweighted categorical cross-entropy CE is considered as objective function. To prevent overfitting on all layers, gradients are back-propagated using an Adam optimizer and a $\mathbb{L}2$-regularization with a weight decay of 1.10^{-6}.

5. Experiments

5.1. The Réunion satellite image time series

A Satellite Image Time Series (SITS) covering the *Réunion* island is used as a baseline in this paper. Its ground truth is available in [37] and deep land cover classifiers reaching relevant performance levels on this dataset have been proposed in [20], [15] and [7]. In more details, this SITS consists in 21 Sentinel-2 images acquired between January and December 2017 and covering a 67 km x 59 km scene with a spatial resolution of 10 metres, i.e. each image contains $6667 \times 5916 = 39441972$ pixels. Channels available at a 10-meter resolution are B2 (blue), B3 (green), B4 (red) and B8 (near-infrared). In addition, the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) [38] are computed and supplied for the same spatial resolution. These standard indexes are defined by $NDVI = f(B8, B4)$ and $NDWI = f(B3, B8)$ with $f(x, y) = \frac{x-y}{x+y}$, a homogeneous function from $\mathbb{R}_+^* \times \mathbb{R}_+^*$ to $[-1; 1]$. Cloud removal is simply performed for each band by linear interpolation of the cloudy pixels based on the previous and subsequent cloud-free acquisitions. For more details, the reader is referred to [39]. Finally, the ground truth, which is available as a set of non-contiguous dispersed polygons, accounts for 2% (880,828 pixels) of the pixels which are annotated according to 11 unbalanced land cover classes. They are listed in Table 2 along with their class ratios and the colour they are associated with when visualized. Figure 2 depicts the ground truth by overlaying land cover classes with a Google Earth Engine (GEE) [40] background image. Using the same technique, Figure 3 and Figure 4 zoom in on the ground truth of two areas that are studied further.

Figure 2: Visualisation of the *Réunion* SITS ground truth: a set of dispersed and non contiguous polygons (880,828 pixels).

Figure 3: The *Réunion* SITS ground truth over the northern part of la Plaine des Cafres. Large polygons of class pasture are present.

Layer	Operation	Specifications	Output Tensor Shape
①	split	C channels	$(T, 3, 3)$ (xC)
② (xC)	conv3d	kernel = $(k_1 = T/3, 3, 3)$, no padding, $n_{filter} = N_1 = 256$	$(T_1, 1, 1, N_1)$
③ (xC)	conv1d	kernel = $(k_2 = T/3)$, no padding, $n_{filter} = N_2 = 64$	(T_2, N_2)
④	stack	C channels	(C, T_2, N_2)
⑤	attention	C weights α_i	
⑥	hidden dense	+ ReLu	256 neurons
⑦	final dense	+ softmax	one neuron / class

Table 1: *X-DSLCC* architecture. Features are extracted by 2 different convolution layers computed for each one of the C channels (C times layers ②, ③). They are stacked with layer ④ whose neuron activations are further considered to build redescriptions. Attention layer ⑤ is outing C channel attention weights.

Class	Ratio (%)	Color
Sugar cane	12.4	
Pasture	7.3	
Market gardening	2.3	
Greenhouse crops	0.2	
Orchards	3.9	
Wooded areas	23.5	
Moor	16	
Rocks	21.5	
Relief shadows	5.1	
Water	6.1	
Urban area	1.8	

Table 2: The *Réunion* SITS: ground truth classes.

Figure 4: The *Réunion* SITS ground truth over the airport of Saint Denis. Polygons of the classes urban, moor, wooded areas and rocks are visible.

5.2. Experimental settings

5.2.1. Network hyperparameters and training partitions

The hyperparameters of *X-DSLCC* (see Section 4.2) are set according to dataset characteristics. Regarding the *Réunion* SITS, since it contains $C = 6$ channels and $T = 21$ acquisitions, the temporal convolution dimension is set to $k_1 = k_2 = T/3 = 7$. As a result, the shape of the output tensors of layers ①, ②, ③ and ④ are, respectively, $(21, 3, 3)$, $(15, 1, 1, 256)$, $(9, 64)$, and $(6, 9, 64)$. The classified pixels being described by the neuron activations at layer ④, 576 activations are made available for each one of the 6 channels.

TensorFlow2 is used to optimize the model. Annotated pixels are divided into a training dataset (60%), a validation (20%) dataset and a test (20%) dataset using a stratified sampling that maintains class ratios. Pixels belonging to a same polygon all belong to the same dataset.

5.2.2. GFS-pattern mining parameters

The free prototype *DFTS-P2miner* [41], whose sources can be downloaded from `https://sites.google.com/view/dfts-p2miner`, is used to run the entire extraction process for each channel, from pixel values quantization to pattern selection. Pixel values are converted to symbols '1', '2', and '3', that respectively report 'low values', 'medium values' and 'high values'. This quantization is detailed in Section 3.1 and commonly adopted to mine GFS-patterns [33]. Greenhouse crops, the class with the lowest representation, has only 1,931 pixels. The minimum surface threshold σ is thus set to 881 pixels under the assumption that about half of these pixels have the same kind of evolution. This is a fairly lax constraint because it only accounts for 0.1% of annotated pixels. Finally, the 120 most interesting maximal GFS-patterns are selected. Each pixel, for a specific channel, is thus characterized by stating whether or not each one of these 120 patterns is present. This choice is conservative as a set of only 40 patterns is recommended in [34]. By doing so, the chance

of matching neuron activations with dataset regularities is increased. All other DFTS-P2miner parameters resort to default values.

5.2.3. Resdecription mining parameters

Using the free prototype *Siren* [42, 43], whose sources are available at `https://gitlab.inria.fr/egalbrun/siren`, redescription mining is performed for each class of interest and each channel. The maximum number of neuron activation variables and pattern variables is respectively set at 4 and 1 to extract expressive yet straightforward redescriptions. Beside providing easy-to-read redescriptions, these syntactic limitations also ensure a reasonable consumption of resources. To extract as many redescriptions as possible, the minimum Jaccard index is set to the Siren default value, 1%. Following this strategy, the minimal number of pixels supporting a redescription is arbitrarily set to 1% and the standard value of 5% is used as the maximum p-value. All other Siren parameters resort to default values.

5.3. Quantitative results

5.3.1. Resource consumption

All experiments were performed on a standard computing platform (AMD Ryzen ThreadRipper 3990X, 4.3 GHz, 256 GB RAM, GeForce RTX 2060 Super) running Linux (kernel 5.11.0-40). The *PM4X* method (see Section 4) was assessed by executing a GPU version of TensorFlow2 (step 1), a single-threaded implementation of *DFTS-P2miner* (step 1), a single-threaded version of *Siren* (step 3) and dedicated Python scripts (steps 2, 4 and 5). The most resource-intensive operations are mining steps 1 and 3. The corresponding execution times and maximum memory consumptions are therefore made available in Table 3. Note that even when considering a single-threaded implementation, the GFS-pattern and redescription mining steps can be fully parallelized across channels, i.e., one execution per channel, with all executions running simultaneously.

5.3.2. Network performances

It is recalled that we are not interested in performance as such, and that *X-DSLCC* is primarily designed to test whether it is possible to correlate neural activation levels with the presence of spatiotemporal patterns in input data. However, in order to illustrate the performance penalty of designing a network that extracts features resembling GFS patterns, the performance of *X-DSLCC* is compared with that of three land cover classifiers that also work at the pixel level and mobilize all available channels. More precisely, a baseline is provided with the performance of a classical random forest (500 trees, 200 splits). The performances of the two inspiring deep neural networks that led to the *X-DSLCC* architecture (see section 4.2), namely *TempCNN* [1] and *Sdeep-B-Multi-ii* [15], are also reported. These network were chosen because, to our knowledge, they deliver the best performances for the Réunion SITS [15] and they work in a similar way to *X-DSLCC*. More precisely, both *TempCNN* and *Sdeep-B-Multi-ii* rely on temporal convolutions which are applied to each channel separately, before merging obtained features for the final decision stage. The accuracy rates for *Random Forest*, *TempCNN*, *Sdeep-B-Multi-ii*, and *X-DSLCC* are 90.4%, 91.3%, 92.2%, and **84.9**% respectively. Although not the best network, *X-DSLCC* reaches a decent accuracy level. Its precision and recall measures are given in Table 4. In the field of land cover classification, according to [44], *X-DSLCC* can be considered insufficiently accurate, as most deep learning-based proposals achieve overall accuracies above 90%. However, when it comes to understanding how a network works, explaining the wrong decisions is just as important as explaining the right ones, which justifies the use of *X-DSLCC* in this paper.

Operation	**Execution time (s)**	**Memory usage (MB)**
Step 1		
LCC (*X-DSLCC*) learning	7 320	47 409
GFS-pattern mining - B2	95 647	3 839
GFS-pattern mining - B3	139 234	3 712
GFS-pattern mining - B4	92 974	3 741
GFS-pattern mining - B8	37 867	4 087
GFS-pattern mining - NDVI	153 346	3 840
GFS-pattern mining - NDWI	87 882	3 828
Step 3		
redescription mining - B2	1 435 579	23 353
redescription mining - B3	1 314 548	23 205
redescription mining - B4	1 524 379	23 375
redescription mining - B8	1 288 256	23 547
redescription mining - NDVI	1 226 986	23 027
redescription mining - NDWI	1 696 710	24 109

Table 3: Resource consumption: mining steps 1 and 3.

Class	**Precision**	**Recall**	**Ratio**
Sugar cane	87.7	91.2	12.4
Pasture	86.9	85.3	7.30
Market gardening	59.5	63.5	2.30
Greenhouse crops	26.5	20.3	0.20
Orchards	59.0	63.4	3.90
Wooded areas	85.3	85.8	23.5
Moor	85.1	79.2	16.0
Rocks	92.3	94.1	21.4
Relief shadows	81.8	91.3	5.10
Water	95.1	82.5	6.10
Urban area	75.0	81.6	1.80

Table 4: Precision and recall by class for *X-DSLCC* on the test set composed of 176,166 pixels samples.

5.3.3. Channel attention weights statistics

The spread and location of channel attention weights over the whole dataset are provided for each class and each channel with Table 5 using 25^{th} percentiles (denoted P25), 50^{th} percentiles (i.e. medians), and 75^{th} percentiles (denoted P75). As it can be observed, channel attention weights differ from one channel to another according to classes. For example, in keeping with median attention weights, the most important channels are:

- B2 for classes sugar cane, pasture, market gardening, orchards, wooded areas and moor,
- B3 for reliefs shadows,
- B4 for water,
- B8 for green house crops, rocks and urban area.

It is important to balance these results with the fact that all bands are exploited by the network, and sometimes with weights that are very close to reported maxima. For example, for classes pasture, market gardening, orchards, wooded areas and moor, channel B2 is the most mobilized one, which is counterintuitive. Indeed, vegetation is generally detected using B4, B8 or NDVI. If we now have a closer look at attention weights, it appears that, for all of these classes, band B4 is also exploited at very high levels that are very close to B2 ones. In other words, the network does rely on expected channels such as B4, but it can also select unexpected bands such as B2. Interestingly, synthetic channels, i.e., vegetation and water indices, are never reported as being the most important channels. Additionally, high attention weights can be applied to either low or high neural output values, i.e., they can amplify any kind of signals, inhibited or strong ones. This variability shows that making conclusions from the sole attention weight values is not precise enough. As explained further in Section 5.5, the B2 features involved in class Pasture redescriptions are mostly inhibited: the attention operator thus exploits this setting, which is an expected behavior for vegetation. Further, one may expect, for a given class, the existence of cohorts, here ground area clusters, with specific behaviours that may lead to different attention values. Redescription mining is also expected to detect and provide insight.

5.3.4. Redescriptions statistics

As explained in Section 4, redescriptions are extracted for each class and each channel separately. Table 6 thus reports, for each class and each channel:

- r, the number of extracted redescriptions ,
- S_{all}, the surface covered by the r redescriptions, i.e. the sum of their supports since they are spatially complementary (cf. Step 4 in Section 4),
- S_{min}, the minimum support, i.e. the minimum surface, observed for the r redescriptions,
- S_{max}, the maximum support, i.e. the maximum surface, observed for the r redescriptions,
- J_{min}, the minimum Jaccard index, i.e. the minimum accuracy, observed for the r redescriptions,
- J_{max}, the maximum Jaccard index, i.e. the maximum accuracy, observed for the r redescriptions.

The median and the standard deviation of these measures are made available at the bottom of the same table.

The reported number of extracted redescriptions is between 0 and 10, which is a reasonable amount of information that can be processed by end users. Their support ranges from 1%, i.e. the minimum number of pixels set to mine them, to 78 %. Their accuracy start from 2.1 %, i.e. twice the minimum accuracy used to extract them, and reach 96 % at the most. The surface covered by the redescriptions for each class and each channel, S_{all}, is between 0% and 90%. Extracted redescriptions thus vary widely in terms of number, support and accuracy, which is confirmed by the standard deviations reported for each measure, especially when compared with their medians. Taking into account the redescriptions extracted from all channels, the Redescription Decision Cover RDC, i.e., the fraction of right decisions for which one or more redescriptions hold is given for each class by Table 7. Results are very encouraging with a median value of RDC that reaches 85.5 %. Furthermore, as regards the performance of the *X-DSLCC* network, the worst decision cover is obtained for greenhouse crops and the best for water. In other words, as can be expected for a data-driven model, there are few redescription occurrences associated with the right functioning of the network when the latter struggles to reach high performances due to the very low-class ratio (0.2 %).

5.4. Qualitative results

Explanations and interpretations for the *X-DSLCC* network are provided in the form of attention weights and redescriptions. This section focuses on two classes of interest, namely pasture and urban area. For the sake of clarity, in the following, maps are generated using the ground truth polygons of a single class, the one that is under consideration.

5.4.1. Attention weights-based explanations

Figure 5 shows the attention weights obtained for class pasture and class urban area using box plots computed for the whole dataset. Such visualizations are useful for quickly identifying at the class level which channels are important in the final decision. In the case of the *Réunion* SITS, channel B2 is predominant for pasture class since that it can show vegetation senescence [45]. Channel B4, which detects maximum chlorophyll absorption [45], is the second most used channel used to detect pasture category. Again, the *X-DSLCC* network focuses on appropriate features. Channel B8 is reported to be the most important for

Class	Channel	P25	Median	P75
Sugar cane	B2	0.91	0.96	0.98
	B3	0.32	0.45	0.56
	B4	0.86	0.91	0.94
	B8	0.73	0.74	0.78
	NDVI	0.65	0.70	0.76
	NDWI	0.84	0.88	0.91
Pasture	B2	0.94	0.96	0.97
	B3	0.34	0.45	0.56
	B4	0.92	0.95	0.96
	B8	0.75	0.78	0.81
	NDVI	0.72	0.75	0.77
	NDWI	0.76	0.83	0.88
Market gardening	B2	0.96	0.98	0.98
	B3	0.53	0.69	0.82
	B4	0.93	0.95	0.97
	B8	0.79	0.83	0.89
	NDVI	0.72	0.75	0.78
	NDWI	0.84	0.88	0.91
Greenhouse crops	B2	0.58	0.70	0.81
	B3	0.04	0.07	0.16
	B4	0.70	0.81	0.88
	B8	0.79	0.82	0.85
	NDVI	0.70	0.78	0.85
	NDWI	0.83	0.86	0.89
Orchards	B2	0.98	0.99	0.99
	B3	0.70	0.81	0.88
	B4	0.96	0.98	0.98
	B8	0.76	0.79	0.83
	NDVI	0.75	0.76	0.78
	NDWI	0.84	0.87	0.90
Wooded areas	B2	0.99	0.99	0.99
	B3	0.88	0.94	0.98
	B4	0.98	0.99	0.99
	B8	0.79	0.84	0.89
	NDVI	0.75	0.77	0.78
	NDWI	0.89	0.90	0.92
Moor	B2	0.96	0.98	0.99
	B3	0.55	0.87	0.95
	B4	0.94	0.97	0.98
	B8	0.81	0.89	0.96
	NDVI	0.73	0.76	0.79
	NDWI	0.75	0.86	0.90
Rocks	B2	0.88	0.97	0.99
	B3	0.39	0.92	0.98
	B4	0.89	0.96	0.98
	B8	0.93	0.98	0.99
	NDVI	0.79	0.88	0.95
	NDWI	0.54	0.80	0.92
Relief shadows	B2	0.98	0.99	0.99
	B3	0.98	0.99	1 .00
	B4	0.98	0.99	0.99
	B8	0.99	0.99	0.99
	NDVI	0.75	0.79	0.84
	NDWI	0.73	0.85	0.91
Water	B2	0.93	0.96	0.98
	B3	0.86	0.93	0.99
	B4	0.97	0.98	0.99
	B8	0.75	0.84	0.93
	NDVI	0.66	0.80	0.89
	NDWI	0.29	0.63	0.83
Urban area	B2	0.68	0.85	0.90
	B3	0.13	0.34	0.55
	B4	0.78	0.88	0.92
	B8	0.87	0.94	0.97
	NDVI	0.73	0.89	0.96
	NDWI	0.59	0.79	0.90

Table 5: Channel attention weights: spreads and locations for each class and each channel.

Class	Channel	r	S_{all}	S_{min}	S_{max}	J_{min}	J_{max}
Sugar cane	B2	4	0.26	0.01	0.20	0.08	0.42
	B3	3	0.40	0.01	0.32	0.36	0.60
	B4	4	0.29	0.02	0.20	0.41	0.66
	B8	3	0.56	0.01	0.45	0.27	0.77
	NDVI	7	0.12	0.01	0.05	0.40	0.71
	NDWI	10	0.24	0.01	0.06	0.46	0.66
Pasture	B2	1	0.23	0.23	0.23	0.47	0.47
	B3	2	0.26	0.07	0.19	0.30	0.43
	B4	2	0.21	0.01	0.20	0.26	0.45
	B8	4	0.42	0.01	0.34	0.14	0.63
	NDVI	2	0.02	0.01	0.01	0.30	0.31
	NDWI	3	0.45	0.02	0.29	0.37	0.68
Market Gardening	B2	3	0.13	0.01	0.10	0.37	0.39
	B3	2	0.21	0.03	0.18	0.35	0.42
	B4	2	0.20	0.01	0.19	0.22	0.52
	B8	5	0.32	0.01	0.18	0.50	0.66
	NDVI	6	0.10	0.01	0.03	0.22	0.82
	NDWI	3	0.42	0.02	0.24	0.42	0.80
Greenhouse crops	B2	4	0.06	0.01	0.02	0.38	0.71
	B3	5	0.06	0.01	0.02	0.25	0.62
	B4	4	0.09	0.01	0.06	0.27	0.69
	B8	4	0.31	0.01	0.14	0.24	0.58
	NDVI	4	0.33	0.01	0.30	0.29	0.92
	NDWI	3	0.36	0.01	0.34	0.53	0.63
Orchads	B2	2	0.21	0.11	0.11	0.34	0.40
	B3	3	0.09	0.02	0.06	0.24	0.36
	B4	2	0.28	0.12	0.16	0.49	0.51
	B8	4	0.51	0.01	0.32	0.40	0.75
	NDVI	2	0.18	0.01	0.17	0.23	0.74
	NDWI	4	0.50	0.01	0.28	0.50	0.79
Wooded areas	B2	3	0.14	0.02	0.08	0.30	0.54
	B3	4	0.12	0.01	0.07	0.12	0.40
	B4	4	0.23	0.01	0.18	0.12	0.42
	B8	3	0.37	0.04	0.20	0.66	0.68
	NDVI	4	0.25	0.01	0.18	0.29	0.72
	NDWI	4	0.60	0.01	0.29	0.63	0.77
Moor	B2	5	0.21	0.01	0.08	0.28	0.40
	B3	4	0.19	0.01	0.14	0.24	0.45
	B4	2	0.18	0.01	0.17	0.32	0.57
	B8	5	0.55	0.05	0.16	0.56	0.82
	NDVI	5	0.11	0.01	0.04	0.33	0.77
	NDWI	5	0.64	0.02	0.30	0.55	0.88
Rocks	B2	5	0.17	0.01	0.07	0.33	0.73
	B3	4	0.15	0.01	0.09	0.07	0.45
	B4	2	0.17	0.06	0.11	0.44	0.51
	B8	5	0.66	0.01	0.54	0.53	0.90
	NDVI	2	0.65	0.06	0.59	0.55	0.79
	NDWI	2	0.60	0.11	0.49	0.64	0.78
Relief shadow	B2	3	0.12	0.01	0.10	0.43	0.87
	B3	4	0.25	0.01	0.17	0.42	0.86
	B4	3	0.07	0.01	0.03	0.10	0.57
	B8	3	0.34	0.01	0.20	0.58	0.68
	NDVI	4	0.25	0.01	0.17	0.41	0.78
	NDWI	4	0.30	0.02	0.10	0.38	0.50
Water	B2	3	0.18	0.01	0.14	0.54	0.69
	B3	8	0.27	0.01	0.17	0.34	0.72
	B4	4	0.16	0.03	0.07	0.48	0.66
	B8	5	0.90	0.02	0.78	0.76	0.96
	NDVI	0	0.00	-	-	-	-
	NDWI	2	0.57	0.01	0.56	0.31	0.70
Urban area	B2	3	0.17	0.01	0.15	0.27	0.66
	B3	3	0.04	0.01	0.02	0.02	0.28
	B4	1	0.04	0.04	0.04	0.57	0.57
	B8	5	0.49	0.03	0.21	0.46	0.74
	NDVI	2	0.69	0.07	0.62	0.53	0.89
	NDWI	3	0.64	0.04	0.47	0.41	0.79
Median		3.50	0.25	0.01	0.17	0.37	0.66
Standard deviation		1.61	0.20	0.04	0.16	0.15	0.17

Table 6: Redescriptions: supports and accuracy for each class and each channel.

Class	*RDC* (%)
Sugar cane	89.3
Pasture	82.7
Market gardening	81.1
Greenhouse crops	63.8
Orchards	84.5
Wooded areas	85.5
Moor	90.2
Rocks	94.7
Relief shadows	79.8
Water	95.8
Urban area	92.5

Table 7: Redescription Decision Cover *RDC* by class.

the detection of urban areas. Since it indicates the presence of biomass [45], which is minimal in the case of urban areas, the *X-DSLCC* network focuses on low B8 values to exhibit them, which is further confirmed by redescriptions. Whatever the class that is considered, pasture or urban area, all channels are exploited at quite high levels, except channel B3 which is sensitive to the total chlorophyll in vegetation [45]. It is thus assumed that similar information is obtained through other channels, especially B4.

These attention-based class-level explanations can be provided together with pixel-level explanations using attention maps. They are created for the pixels of a given class and a given channel by 1) normalising attention weights between 0 and 255, and 2) depicting them using a colour scale whose dominant colour matches as closely as possible the radiometry of the channel under consideration. Pixels not identified as belonging to the class of interest are simply filled with natural colours. This overlay is performed using Google Earth Engine (GEE) [40]. Figure 6 and Figure 7 give an example of such a map for class pasture and class urban area using channel B2 and channel B8 respectively. These maps show that even though selected channels are dominant at the class level, they are not systematically focused on at the pixel level. In addition, low weights do not necessarily imply wrong decisions, and high weights do not guarantee right decisions: several attention weight settings are learnt by *X-DSLCC* to predict land cover classes. These settings can be conveniently rendered for each pixel using a histogram showing the attention weights across all bands and used to predict its class. Such a histogram can be for example displayed when hovering over a pixel of interest in a geographical information system.

5.4.2. Redescriptions-based explanations

They are supplied with their metrics (support, Jaccard index) and visualized thanks to their spatiotemporal maps (see Section 1) to refine explanations conveyed by attention weights. Each map illustrates where a re-

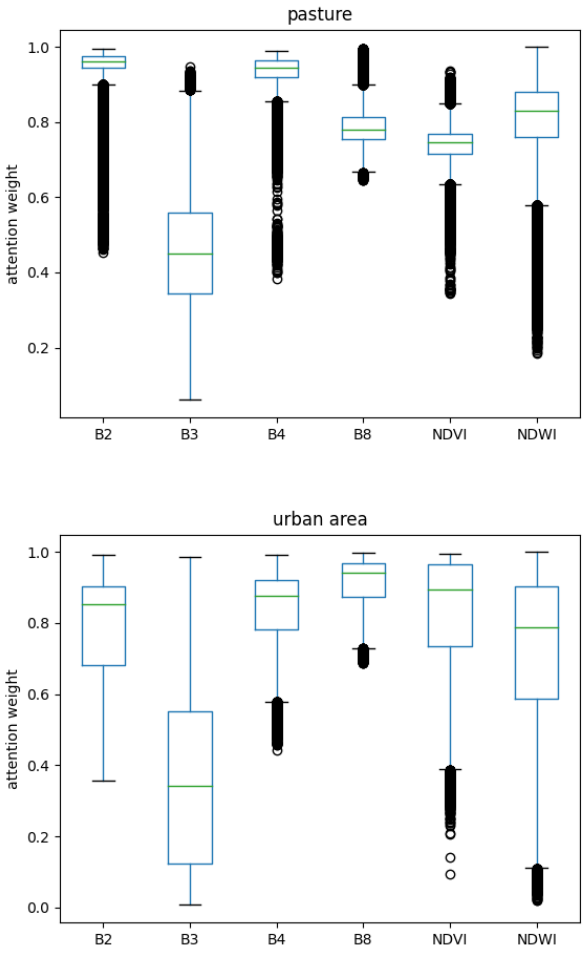

Figure 5: Channel attention weights: box plots for classes pasture and urban area.

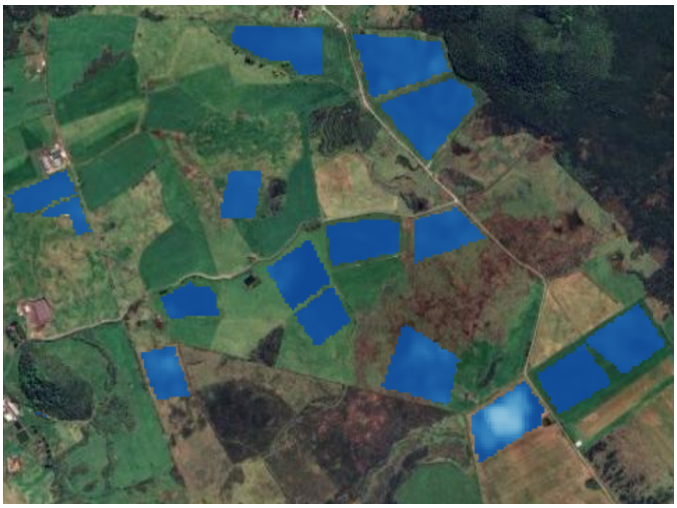

Figure 6: Attention map for class pasture and channel B2 (blue colour scale) over the northern part of la Plaine des Cafres.

Figure 7: Attention map for class urban area and channel B8 (red colour scale) over the airport of Saint-Denis.

description holds in space and in time. The ending date of the occurrences of the redescription pattern is denoted using the temporal colour palette depicted by Figure 8. Pixels that are no explained by the redescription are colored as follows: the white colour is used to indicate decisions where the activation-level conditions are not met and the pattern is absent, i.e. the redescription does not hold. The brown colour is associated with decisions where the activation-level conditions are met and the pattern is absent. If the pattern is present and the activation-level conditions are not met, the grey colour is assigned to the pixels. **These settings are listed in Table 8**. The black colour is reserved for false negatives. Other pixels are filled with the colours of a satellite acquisition, once again using GEE [40]. For the sake of clarity, only redescriptions whose support is greater than or equal to 10% will be discussed.

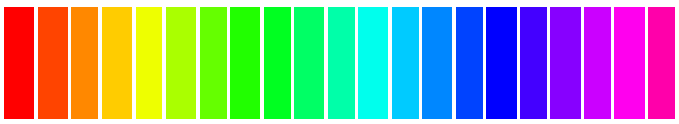

Figure 8: Temporal colour palette for covered pixels: 21 acquisitions, from January 2017 (red) to December 2017 (magenta).

Meaning	Color
Activation levels not met, pattern absent	
Activation levels not met, pattern present	
Activation levels met, pattern asbent	

Table 8: Uncovered pixels color palette.

Class Pasture: right decisions

The redescritpions and maps of class pasture are given with figures 9, 10, 11, 12, 13 and 14 for pastures located in the northern part of la Plaine des Cafres. As a reminder, interpretations of how the network works are provided with

Figure 9: B2: $r0: \ -1.665398 < a_{353} < -0.259085 \wedge 0.280056 < a_{538} < 1.904732 \wedge a_{558} < 0.471004 \wedge -0.981744 < a_{569} < -0.030657 \sim 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2$ ($S = 0.22, \ J = 0.47$) over true positives. False negatives appear in black.

the left-hand sides of the redescriptions. They indicate the activation levels of the neurons that are assumed to represent the right-hand sides. The latter, considered as explanations, are expressed as temporal patterns consisting of the symbols '1', '2' and '3', denoting respectively 'low', 'medium' and 'high' reflectance values observed for the channel from which they are extracted. Redescriptions $r0$ and $r1$ (figures 9 and 10), extracted from channels B2 and B3, can be related to vegetation senescence and drying (rainfall decreases during the summer). Indeed, their patterns contain series of high reflectance values (symbols '3') followed by series of intermediate values (symbols '2'). According to their maps, the occurrences of the $r0$ and $r1$ patterns end in late 2017 (dark blue, violet and magenta colours) and are spatially quite complementary. Redescription $r2$ (Figure 11), extracted from channel B4, expresses a similar phenomenon, but with two differences: 1) it first traces an increase in chlorophyll presence before showing a clear decrease, and 2) it tends to appear earlier in the series (green and light blue colours). In addition, a continuous presence of biomass at high levels is evidenced for channel B8 with redescription $r3$ (Figure 12), while redescriptions $r4$ and $r5$ (Figure 13 and Figure 14) respectively show a continuous plant water content at medium levels and an increase of the latter. Once again, these maps are quite complementary spatially. Finally, it is recalled that all of these explanations refine attention-based ones. For example, relying on channels B2 and B4, the two most used channels according to the channel-based attention operator (see Figure 5), to identify class pasture is consistent as long as decreasing reflectance values are present in the input data, which is evidenced by redescriptions $r0$ and $r2$.

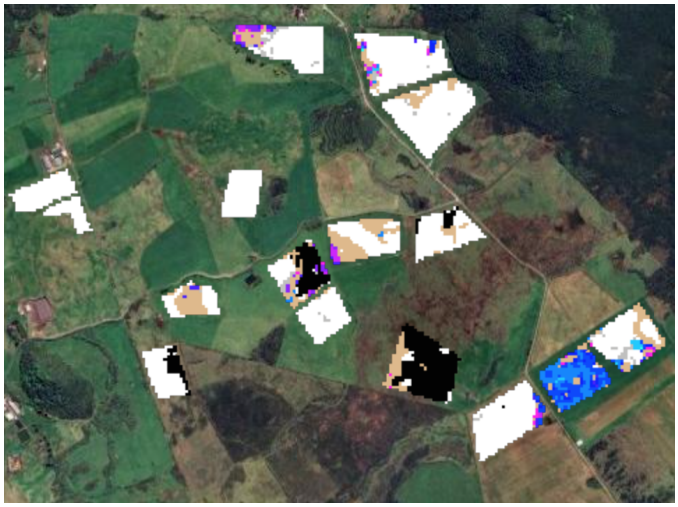

Figure 10: B3: $r1: \ -0.628419 < a_{139} < 0.463561 \wedge -0.824907 < a_{267} < -0.343454 \wedge 1.005415 < a_{294} < 1.641748 \wedge -0.748222 < a_{523} < -0.351193 \sim 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$ $(S = 0.19, \ J = 0.43)$ over true positives. False negatives appear in black.

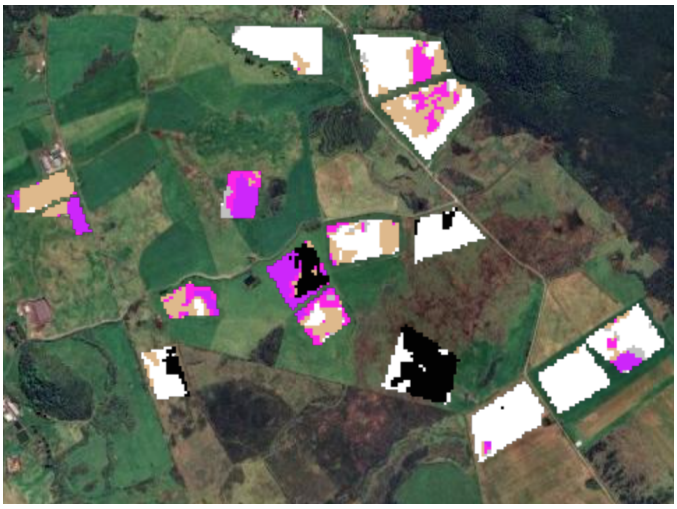

Figure 12: B8: $r3: \ -7.776212 < a_{24} < -3.750202 \wedge -2.905304 < a_{164} < -1.271326 \wedge -3.461838 < a_{522} < -1.903169 \wedge -3.75082 < a_{548} < -1.395322 \sim 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3$ $(S = 0.34, \ J = 0.63)$ over true positives. False negatives appear in black.

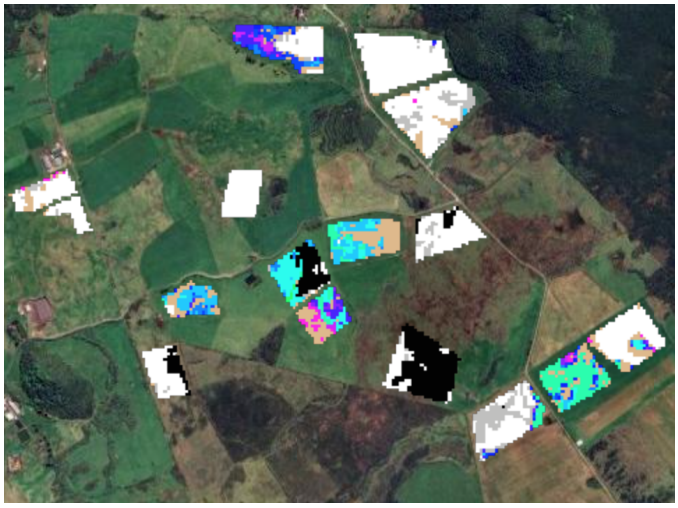

Figure 11: B4: $r2: \ 0.666302 < a_{53} < 1.775739 \wedge -1.932449 < a_{59} < -0.963242 \wedge -0.571004 < a_{529} < 0.856778 \wedge -1.584851 < a_{560} < -0.513726 \sim 2 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2$ $(S = 0.2, \ J = 0.45)$ over true positives. False negatives appear in black.

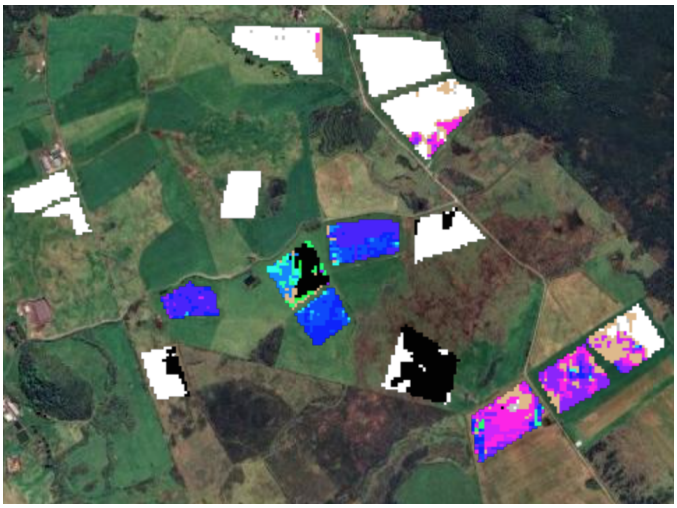

Figure 13: NDWI: $r4: \ -6.364172 < a_{60} < -1.538447 \wedge 0.765226 < a_{77} < 5.49521 \wedge 3.773017 < a_{547} < 8.465077 \wedge a_{553} < -0.301693 \sim 2 \to 2 \to 2 \to 2 \to 2 \to 3 \to 3 \to 3 \to 3 \to 3$ $(S = 0.3, \ J = 0.68)$ over true positives. False negatives appear in black.

Figure 14: NDWI: $r5: \ 1.030979 < a_{28} < 2.855336 \wedge 0.381648 < a_{226} < 3.289396 \wedge 0.469679 < a_{481} < 3.046999 \wedge -2.07497 < a_{553} < 0.569858 \sim 2 \to 2$ ($S = 0.14, \ J = 0.55$) over true positives. False negatives appear in black.

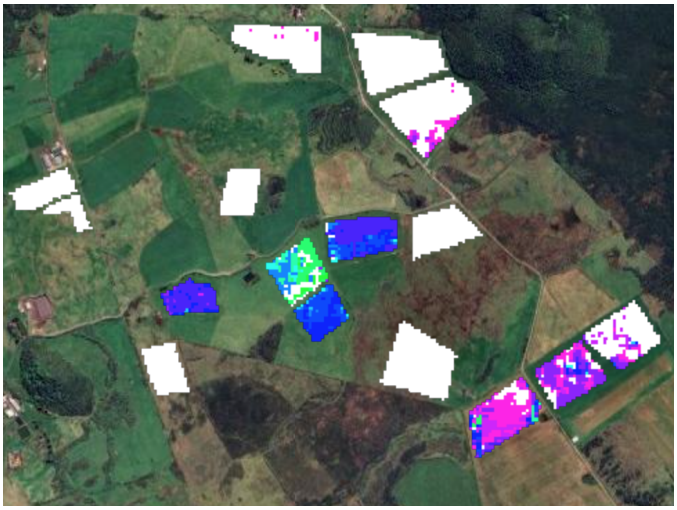

Figure 15: NDWI: $r4$ pattern $2 \to 2 \to 2 \to 2 \to 2 \to 3 \to 3 \to 3 \to 3$ over the ground truth.

Class Pasture: wrong decisions

With regard to the false negatives, i.e. the black pixels observed in the previous maps, it is possible to visualize the patterns of redescriptions over the ground truth, rather than the correct decisions as done previously. The temporal palette used for the previous maps is used once again, and pixels not affected by the patterns are simply set to white. Such visualizations allow checking whether redescription patterns are present in the dataset and should have been captured by the network or not. As for the pasture in the centre of the image, it appears that the patterns of redescriptions $r1$, $r2$, $r3$ and $r4$ are present for most of the false negatives. An example of such a map is given with Figure 15 for $r4$. The network and/or the learning dataset should thus be revised to detect these patterns and classify these false negatives correctly. The nature and the number of the features learnt for each channel, and the fusion between these features, could be for example questioned. The learning dataset could also be augmented by generating new samples containing the patterns identified by redescriptions. Another analysis can be conducted to identify which pasture patterns should not be considered by the network in the case of false positives. For example, moor can be confused with pasture category, and, among the patterns that could explain this mistake, the $r5$ one is the most present for the area shown in Figure 16. In this figure, the map of redecription of $r5$ is made available for false positives. In addition, the green-khaki colour mark areas were moor category was correctly identified by the network.

Figure 16: NDWI: $r5: 1.030979 < a_{28} < 2.855336 \wedge 0.381648 < a_{226} < 3.289396 \wedge 0.469679 < a_{481} < 3.046999 \wedge -2.07497 < a_{553} < 0.569858 \sim 2 \to 2$ ($S = 0.14, \ J = 0.55$) over the false positives. Green khaki: moor, correctly identified by the network.

Figure 17: B2: $r6: \ -1.923968 < a_{84} < -0.942249 \wedge 0.652257 < a_{366} < 1.600037 \wedge -1.296196 < a_{523} < -0.824066 \wedge 0.455249 < a_{558} < 0.960992 \sim 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2$ ($S = 0.15, \ J = 0.51$) over true positives. False negatives appear in black.

Figure 18: B8: $r7: \ -0.916848 < a_{24} < 0.030338 \wedge -0.503101 < a_{164} < -0.109455 \wedge -0.67026 < a_{420} < -0.030818 \wedge a_{486} < 1.474579 \sim 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$ ($S = 0.2, \ J = 0.74$) over true positives. False negatives appear in black.

Class Urban Area: right decisions

Four redescriptions and their maps are provided with figures 17, 18, 19, and 20. For the considered scene, the ground truth includes the airport, some buildings to the east of the airport and a stretch of road to the southwest. Redescription $r6$ shows a drop from high levels in channel B2 to medium levels, which may be related to the presence of dark asphalt surfaces. Redescriptions $r7$, $r8$ and $r9$ capture a weak and continuous presence of vegetation, biomass and water respectively, consistent with urbanized areas. These redescriptions confirm for example that focusing on channels B8, NDVI and NDWI, the three most used channels according to the channel-based attention operator (see Figure 5), to infer class urban areas makes sense if low reflectance values are considered. As $r7$, $r8$ and $r9$ are based on patterns containing many symbols, their end dates tend to occur in late 2017, which explains a weak temporal dispersion (most pixels are purple).

Class Urban Area: wrong decisions

Figure 19: NDVI: $r8: \ a_8 < 1.921249 \wedge -1.166227 < a_{534} < 2.407826 \wedge -1.309819 < a_{541} < 1.668957 \wedge a_{554} < 1.277908 \sim 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$ ($S = 0.62, \ J = 0.9$) over true positives. False negatives appear in black.

Figure 20: NDWI: $r9: \ a_{63} < 0.017922 \wedge a_{539} < -0.851809 \wedge 0.246729 < a_{553} < 10.347095 \wedge 0.935444 < a_{570} \sim 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$ ($S = 0.47, \ J = 0.79$) over true positives. False negatives appear in black.

Figure 21: NDVI: $r8$ pattern $1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1$ over the ground truth.

A large part of false negatives could have been identified using the redescriptions patterns, especially the one of $r8$. This can be observed with its map over the ground truth, Figure 21. Nevertheless, $r8$ can lead to false positives and should be balanced with additional information, i.e. other patterns, especially those present in other channels. Figure 22 show these false positives for an area where $r8$ holds and rocks should have been identified. True positives, i.e. rocks, are denoted in light green. Once again, these insights are assumed to guide experts when designing the network and forming the learning dataset.

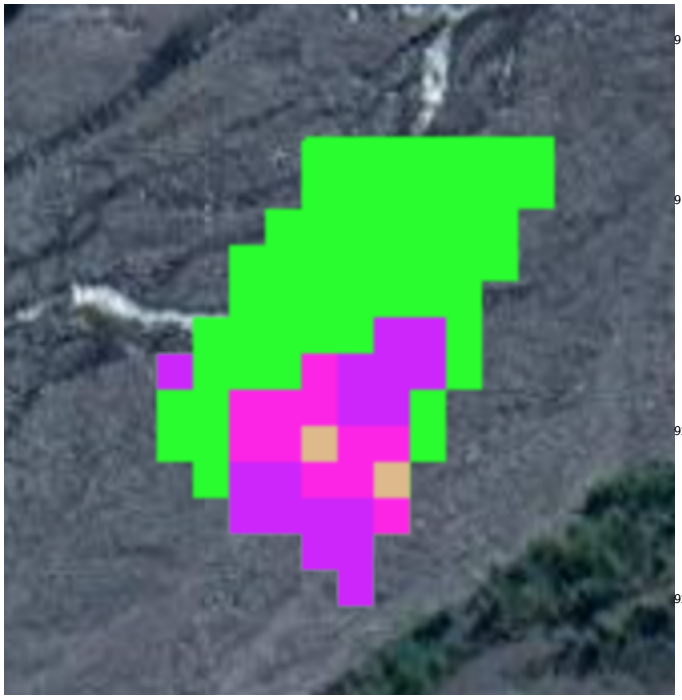

Figure 22: NDVI: $r8: \ a_8 < 1.921249 \wedge -1.166227 < a_{534} < 2.407826 \wedge -1.309819 < a_{541} < 1.668957 \wedge a_{554} < 1.277908 \sim 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \twoheadrightarrow 1 \ (S = 0.62, \ J = 0.9)$ over the false positives. Light green: rocks, correctly identified by the network.

5.5. Redescription-based interpretations

Regarding interpretations, each redescription comes with the neurons implied in the decisions. In addition, their activation levels are supplied as well as their numeros. End-users can thus check whether explanation patterns are captured using a complex setting or not, i.e., whether numerous or few neurons are implied. Activation levels also indicate whether corresponding features are amplified or inhibited by the network. These features can be precisely identified thanks to the numeros of the neurons. Since the latter are located at layer ④ and are simply obtained by stacking the 1D vectors extracted by layer ③, we get $9 * 64 = 576$ neurons for each band, i.e., 9 temporal components for each one of the 64 filters that are learnt. Using the numero of a neuron, it is thus possible to identify the filter it originates from and the temporal component it represents.

For classes Pasture and Urban Area, all redescriptions contains 4 neurons which is the maximum number of neurons set for redescriptions extraction. Their complexity is thus equivalent. According to the redescriptions presented previously, B2, B3, and B4-based features are mostly inhibited for class Pasture with activation levels close to 0, ranging from -1.66 to 1.90. On the contrary, B8 features are negatively amplified, up to level -7.78, and NDWI features are mostly positively amplified, up to level 8.46, which is expected when dealing with vegetation. For class Urban Area, all features tend to be inihbited with activation levels ranging from -1.92 to 2.41, except a NDWI one that is amplified up to a value of 10.38, the highest activation level reported so far. The corresponding pattern expresses a continuous absence of water. In other words, low NDWI levels are amplified to detect urban areas, which is also coherent when processing such a class.

6. Discussion

6.1. On the type of network that can be considered with PM4X

As explained in Section 4.2, *X-DSLCC*, including its input variables, is designed to get, for each channel, 1D tensors that resemble as much as possible GFS-patterns, the latter being also extracted from each channel separately. Redescriptions are thus built from these 1D tensors, once stacked, for each channel separately. Final classification layers are not considered for redescription mining since a single neuron can merge the information of several tensors originating from a same channel and/or different channels. If alternative input variables or model structures were to be considered, then either some features are designed to match GFS-patterns, as was done for *X-DSLCC*, and redescriptions could still be extracted, or no features are intended to be similar to GFS-patterns and redescriptions should not be considered.

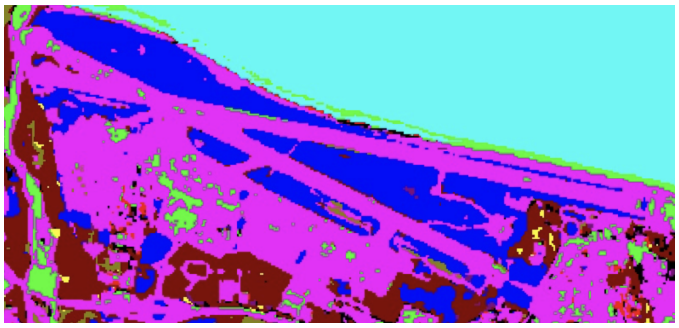

Figure 23: Land cover map produced by *X-DSLCC*. Color legend in Figure 2.

6.2. On classifying unlabelled data

Like any classifier that is assumed to be general, *X-DSLCC* can label undefined areas. Since attention weights are computed at inference time, they can be provided for each pixel. As for redescriptions, since they are determined for a classifier that is general, they are assumed to be general as well. They can therefore be mobilised for interpretations and explanations, which can be done by simply checking whether or not they hold for each pixel, i.e., by verifying whether neural activation level conditions are met and GFS-patterns are present in the input data. Examples of such inferences are generally not provided by LCC papers, since commenting on them without any external, and thus objective reference is questionable. Nevertheless, an outlook is here proposed to show that attention-based and redescription-based explanations are also handy when dealing with unlabelled data. It is based on the Saint Denis airport area as it is a more diverse scene than the one of la Plaine des Cafres. This area and the ground truth polygons it includes are visible in Figure 4. The land cover map produced by *X-DSLCC* is shown by Figure 23. As can be observed, quite homogeneous and plausible zones are exhibited though it seems, without certainty, that some wooded areas are detected as orchards. If class urban area is to be studied, then one can first look at the attention weights observed for band B8, i.e., the band most mobilised by the network to infer this class (see Figure 5). In terms of redescriptions, for band B8, $r7$ has been identified as relevant (see Figure 18). Its map, for all urban class decisions, is shown by Figure 24. The reader is referred to Section 5.4.2 for the interpretation of the colors used to build such a map. As expected, the visualisation remains the same for the areas of the learning data set and is now available for the surrounding unlabelled areas. Figure 24 clearly shows that $r7$ is valid for the airport but not for surrounding urban areas, which is inline with the B8 attention map. On the other hand, redescription $r8$ is valid for both the airport and the surrounding areas, as evidenced by Figure 25. Redescriptions $r7$ and $r8$ all exhibit low level of vegetation, which is consistent with urban areas.

Figure 24: B8: $r7: \ -0.916848 < a_{24} < 0.030338 \wedge -0.503101 < a_{164} < -0.109455 \wedge -0.67026 < a_{420} < -0.030818 \wedge a_{486} < 1.474579 \sim 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$ ($S = 0.2, \ J = 0.74$) for urban area predictions. See Section 5.4.2 for color interpretation.

Figure 25: NDVI: $r8: \ a_8 < 1.921249 \wedge -1.166227 < a_{534} < 2.407826 \wedge -1.309819 < a_{541} < 1.668957 \wedge a_{554} < 1.277908 \sim 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$ ($S = 0.62, \ J = 0.9$) for urban area predictions. See Section 5.4.2 for color interpretation.

Since GFS-patterns are extracted in a fully unsupervised manner, extracting them from unlabelled areas is also possible. Establishing redescriptions for unlabelled areas is thus imaginable by relating activation levels obtained at inference time to the presence of GFS-patterns. In this case, redescriptions would be valid for inferred classes and not actual classes, which could lead to erroneous interpretations and explanations. In addition, such an option is not recommended, because, 1) as stated previously, redescriptions extracted for the ground truth are assumed to be general, and 2) extracting GFS-patterns and redescriptions from unlabelled data would be resource consuming.

6.3. On revising the dataset and the network architecture according to redescriptions

Even tough this proposal is not aimed at revising the dataset and the network in a first place, it is expected that explanations-based insights into false positives and negatives from redescriptions could help to revise achieve such tasks. These should be automated and are still open questions. One direction is to check whether a pattern that is usually detected for a given class is missed out by the network for some pixels. If so, and if it generally ends later than those of the pattern occurrences that are captured by the network for the same class, then this could advocate for convolutional filters whose temporal dimension should be larger. In other words, the current filters may be over-compressing the temporal information.

Going further, knowledge of both the neuron activation levels with respect to the task and the redescription mining results can provide valuable insights for the improvement of the model architecture. A potential application example is the process of model pruning which is commonly applied to enhance inference speed, memory consumption and computational costs by removing neurons and connections for which low activation levels are reported. In such a context, taking into account the redescription results can inform pruning by verifying the co-occurrence of low-activated and non-task-relevant neurons. This also serves as a safety mechanism by signaling the few but relevant enough task-related neuron activations in order to avoid their pruning, which would introduce inference bias, generally affecting under-represented but interesting samples. Further, ensuring that pruning does not degrade model performance could be assessed using permutation-based methods directly applied to the layer for which redescriptions are extracted. Another case study concerns the detection of low model capacity, which can be addressed by adding more neurons and layers. Redescriptions can indeed assist in identifying potentially conflicting rules involving similar neurons and narrow activation levels thus encouraging the expert to increase model capacity. Complementarity with task conflict detection performed in the training step such as done by [46] is therefore worth investigating.

Finally, with regard to the revision of the dataset, a straightforward direction is to point out classes for which the redescription cover is low due to a lack of learning examples or the intraclass variability, thus suggesting to enrich the training dataset appropriately.

6.4. On establishing the temporal preferences of the network

The preference of a network for specific temporal images or phenological characteristics of crops could be assessed by checking all occurrences of all symbols of the redescription patterns. However, this requires a definition of which symbol occurrences should be considered. At present, the pattern occurrence definition only states that the last symbol is the earliest that can be found. All other symbols, as long as they occur earlier and follow the order expressed by the pattern, can occur anywhere in time. Should we look at the earliest ones? The latest ones? The ones in between? Should it depend on the application? This remains an open question. An alternative would be to trace back the temporal components exploited by the network by checking the numero of the redescription neurons. However, since these neurons are located after temporal convolutions, the exact most important temporal components can not be identified precisely. Following the proposal of [35], an efficient last alternative would be to incorporate a temporal attention operator in X-DSLCC, which is a planned evolution of this network.

7. Conclusion

This paper presents an original method for explaining the decisions of a CNN performing LCC based on SITS. This method generates redescription rules, i.e. correspondence rules between neural activation levels and the presence of spatiotemporal patterns in the input data. The activation levels provide interpretations, i.e. information about the functioning of the network, while the spatiotemporal patterns are explanations of the decisions that can be visualized in both time and space. Although redescriptions are extracted for each class, i.e. they are class-level explanations, they can be exploited at the pixel level by checking whether or not they apply to each pixel. Since they are extracted for each channel seperately, these redescriptions can detail the explanations provided by a channel-based attention operator. Experiments on a Sentinel-2 SITS show that such explanations and interpretations can be used to refine channel-attention-based explanations and understand true positives. False positives and negatives can also be assessed by checking whether redescription patterns should have been captured by the network or not. It is anticipated that such an assessment should help end users to revise the learning datasets and the network accordingly. These revision tasks should be automated and form part of our future work directions.

Although up to 95.2% of class decisions can be explained with such a method, full decision coverage is not achieved. Cross-channel redescriptions could therefore be

considered. Furthermore, since the redescriptions are extracted using state-of-the-art incomplete data mining heuristics, it is impossible to ensure that all redescriptions obeying the extraction parameters are found. Besides the incompleteness, and although the syntax of the redescription is limited to very simple expressions (only conjunctions, four activation levels as maximum, one pattern as maximum), these heuristics tend to consume a lot of resources in terms of CPU time and memory. One direction to take to get all the results with fewer resources would be to define activation intervals statically to get only Boolean descriptions and rely on frequent pattern-based redescriptions [35]. However, the definition of such intervals remains an open question. Additionnaly, although attention weights are made available and the spatiotemporal patterns captured by the network are revealed under a p-value constraint, the degree to which the patterns are exploited by the final classification layers is not accessible directly. Finally, the generality of the proposed approach is not demonstrated. The method should be evaluated to see check whether redescriptions provide meaningful explanations regardless of the dataset and the network architecture. Our future work is based on these limitations and also includes improving the performance of *X-DSLCC*while ensuring that features similar to GFS-patterns are extracted.

References

[1] C. Pelletier, G. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sensing 11 (5) (2019) 523. doi:10.3390/rs11050523.

[2] W. Ding, M. Abdel-Basset, H. Hawash, A. M. Ali, Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, Information Sciences (2022).

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[4] R. Xu-Darme, G. Quénot, Z. Chihani, M.-C. Rousset, Sanity checks and improvements for patch visualisation in prototype-based image classification, working paper or preprint (Jan. 2023).

[5] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, S.-M. Hu, Attention mechanisms in computer vision: A survey, Computational Visual Media 8 (3) (2022) 331–368.

[6] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, D. Dou, Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond, Knowledge and Information Systems 64 (12) (2022) 3197–3234.

[7] N. Méger, H. Courteille, A. Benoit, A. Atto, D. Ienco, Explaining a deep spatiotemporal land cover classifier with attention and redescription mining, in: The XXIV International Society for Photogrammetry and Remote Sensing Congress, Vol. XLIII-B3-2022, Nice, France, 2022, pp. 673–680. doi:10.5194/isprs-archives-XLIII-B3-2022-673-2022.

[8] E. Galbrun, P. Miettinen, From black and white to full color: extending redescription mining outside the Boolean world, Statistical Analysis and Data Mining 5 (4) (2012) 284–303. doi:10.1002/sam.11145.

[9] R. Li, S. Zheng, C. Duan, L. Wang, C. Zhang, Land cover classification from remote sensing images based on multi-scale fully convolutional network, Geospatial Information Science 25 (2) (2022) 278–294. doi:10.1080/10095020.2021.2017237.

[10] C. J. Tucker, Red and photographic infrared linear combinations for monitoring vegetation, Remote Sensing of Environment 8 (2) (1979) 127–150. doi:https://doi.org/10.1016/0034-4257(79)90013-0.

[11] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, P. M. Atkinson, Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery, ISPRS Journal of Photogrammetry and Remote Sensing 181 (2021) 84–98. doi:https://doi.org/10.1016/j.isprsjprs.2021.09.005.

[12] D. Ienco, R. Gaetano, C. Dupaquier, P. Maurel, Land cover classification via multitemporal spatial data by deep recurrent neural networks, IEEE Geoscience and Remote Sensing Letters 14 (10) (2017) 1685–1689. doi:10.1109/LGRS.2017.2728698.

[13] M. Rußwurm, M. Körner, Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1496–1504. doi:10.1109/CVPRW.2017.193.

[14] L. Zhong, L. Hu, H. Zhou, Deep learning based multi-temporal crop classification, Remote Sensing of Environment 221 (2019) 430–443. doi:https://doi.org/10.1016/j.rse.2018.11.032.

[15] H. Courteille, A. Benoît, N. Méger, A. M. Atto, D. Ienco, Channel-based attention for land cover classification using sentinel-2 time series, in: IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2021, Brussels, Belgium, July 11-16, 2021, IEEE, 2021, pp. 1077–1080. doi:10.1109/IGARSS47720.2021.9554205.

[16] R. Interdonato, D. Ienco, R. Gaetano, K. Ose, Duplo: A dual view point deep learning architecture for time series classification, ISPRS Journal of Photogrammetry and Remote Sensing 149 (2019) 91–104. doi:https://doi.org/10.1016/j.isprsjprs.2019.01.011.

[17] R. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, D. Lobell, Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods, in: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2019, 2019, pp. 75–82.

[18] M. Rußwurm, M. Körner, Multi-temporal land cover classification with sequential recurrent encoders, ISPRS International Journal of Geo-Information 7 (4) (2018) 129. doi:10.3390/ijgi7040129.

[19] A. Sharma, X. Liu, X. Yang, Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks, Neural Networks 105 (2018) 346–355. doi:https://doi.org/10.1016/j.neunet.2018.05.019.

[20] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, R. Interdonato, Weakly supervised learning for land cover mapping of satellite image time series via attention-based cnn, IEEE Access 8 (2020) 179547–179560. doi:10.1109/ACCESS.2020.3024133.

[21] S. Ji, C. Zhang, A. Xu, Y. Shi, Y. Duan, 3d convolutional neural networks for crop classification with multi-temporal remote sensing images, Remote Sensing 10 (2) (2018) 75. doi:10.3390/rs10010075.

[22] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 193–209.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.

[24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran

Associates, Inc., 2017, pp. 4765–4774.

[25] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: deep learning for interpretable image recognition, Advances in neural information processing systems 32 (2019).

[26] M. Nauta, R. Van Bree, C. Seifert, Neural prototype trees for interpretable fine-grained image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14933–14943.

[27] S. Gautam, M. M.-C. Höhne, S. Hansen, R. Jenssen, M. Kampffmeyer, This looks more like that: Enhancing self-explaining models by prototypical relevance propagation, Pattern Recognition 136 (2023) 109172.

[28] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, P. Watrin, Is attention explanation? an introduction to the debate, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 3889–3900.

[29] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nature Reviews Physics 3 (6) (2021) 422–440.

[30] A. M. Atto, R. R. Bisset, E. Trouvé, Frames learned by prime convolution layers in a deep learning framework, IEEE Transactions on Neural Networks and Learning Systems 32 (7) (2021) 3247–3255. doi:10.1109/TNNLS.2020.3009059.

[31] S. Wang, X. Yu, P. Perdikaris, When and why pinns fail to train: A neural tangent kernel perspective, Journal of Computational Physics 449 (2022) 110768.

[32] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The kdd process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34. doi:10.1145/240455.240464.

[33] A. Julea, N. Méger, P. Bolon, C. Rigotti, M.-P. Doin, C. Lasserre, E. Trouvé, V. N. Lazarescu, Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns, IEEE Transactions on Geoscience and Remote Sensing 49 (4) (2011) pp.1417–1430. doi:10.1109/TGRS.2010.2081372.

[34] N. Méger, C. Rigotti, C. Pothier, T. Nguyen, F. Lodge, L. Gueguen, R. Andréoli, M.-P. Doin, M. Datcu, Ranking evolution maps for Satellite Image Time Series exploration: application to crustal deformation and environmental monitoring, Data Mining and Knowledge Discovery 33 (1) (2019) 131–167. doi:10.1007/s10618-018-0591-9.

[35] E. Galbrun, P. Miettinen, Redescription Mining, SpringerBriefs in Computer Science, Springer, Cham, 2017. doi:10.1007/978-3-319-72889-6.

[36] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd ICML ICLR 2015, 2015, pp. 1–15.

[37] S. Dupuy, R. Gaetano, L. Le Mézo, Mapping land cover on reunion island in 2017 using satellite imagery and geospatial ground data, Data in Brief 28 (2020) 104934. doi:https://doi.org/10.1016/j.dib.2019.104934.

[38] S. K. McFeeters, The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features, International Journal of Remote Sensing 17 (7) (1996) 1425–1432. doi:10.1080/01431169608948714.

[39] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, I. Rodes, Operational high resolution land cover map production at the country scale using satellite image time series, Remote Sensing 9 (2017) 95. doi:10.3390/rs9010095.

[40] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google earth engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment (2017). doi:10.1016/j.rse.2017.06.031.

[41] T. Nguyen, N. Méger, C. Rigotti, C. Pothier, N. Gourmelen, E. Trouvé, A pattern-based mining system for exploring Displacement Field Time Series, in: 19th IEEE International Conference on Data Mining (ICDM) Demo, IEEE, Beijing, China, 2019, pp. 1110–1113.

[42] E. Galbrun, P. Miettinen, Siren: An Interactive Tool for Mining and Visualizing Geospatial Redescriptions , in: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12, Beijing, China, 2012, p. 1544–1547.

[43] E. Galbrun, P. Miettinen, Mining redescriptions with Siren, ACM Transactions on Knowledge Discovery from Data (TKDD) 12 (1) (2018) 1–30. doi:10.1145/3007212.

[44] S. Zhao, K. Tu, S. Ye, H. Tang, Y. Hu, C. Xie, Land use and land cover classification meets deep learning: A review, Sensors 23 (21) (2023). doi:10.3390/s23218966.

[45] Spectral characteristics viewer, `https://landsat.usgs.gov/spectral-characteristics-viewer`, accessed: 2023-05-25 (2023).

[46] S. Guangyuan, Q. Li, W. Zhang, J. Chen, X.-M. Wu, Recon: Reducing conflicting gradients from the root for multi-task learning, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, p. 20.