Grad-SLAM: Explaining Convolutional Autoencoders' Latent Space of Satellite Image Time Series

Thomas Di Martino, Régis Guinvarc'h, Laetitia Thirion-Lefevre, Élise Colin

Grad-SLAM: Explaining Convolutional Autoencoders' Latent Space of Satellite Image Time Series

Thomas Di Martino, Régis Guinvarc'h, Laetitia Thirion-Lefevre, and Élise Colin

***Abstract*—This paper introduces a tool for explaining the latent space generated by applying convolutional autoencoders to satellite image time series, entitled Grad-SLAM. We rely on backpropagated gradient interpretation combined with network activation localization. We use the proposed formula for multiple layers of the encoder, then scale and merge the results to generate a single date contribution metric for the generation of the latent space. We illustrate the potential of this method with the study of the unsupervised classification of agricultural Sentinel-1 time series. We show that critical characterizing dates for unsupervised retrieval of a given class are conditioned by the crop type's radiometric signature and class count. We also present how Grad-SLAM can be used to enhance the understanding of unsupervised classification confusion.**

***Index Terms*—Explainability, neural networks, satellite image time series, unsupervised learning.**

I. Introduction

THE applicability of machine learning algorithms to remote sensing data has been demonstrated by various results [1], [2]. A rising number of these applications rely on autoencoder architectures [3], [4], following the principles and applicative success of representation learning [5]. However, they provide hardly interpretable results [6]. In [7], the authors expose the necessity of explainable machine learning to tackle the black box behavior of such algorithms. Consequently, a growing number of studies involve the application of explainable machine learning, such as in [8], where the authors present a non-black-box model, which raises the curtain on its prediction decisions regarding crop yield prediction from satellite data. Other studies leverage explainable machine learning either for classification [9] or for crop characterization [10].
However, a majority of the explainable machine learning literature is written under the scope of supervised learning. An example of this is the Grad-CAM algorithm [11], which enables the generation of an attention heatmap to highlight regions of an input image responsible for a given classification. Few existing studies involving gradient techniques are used in the context of explainable autoencoders [12], [13], [14].

Building on the aforementioned lines of work, we thus present an adaptation of the Grad-CAM algorithm to explain convolutional autoencoders (CAE) applied to satellite image time series, which we call Grad-SLAM, for Gradient Sequential Latent Activation Mapping. We first introduce convolutional autoencoders, and the explainibility problems intrinsic to their functioning. Then, we present the Grad-CAM approach, and we detail the formula changes involved in this adaptation to CAEs. Then, to illustrate the usage of Grad-SLAM, we set the applicative context of unsupervised crop type classification of Sentinel-1 time series using a CAE. The Grad-SLAM algorithm is thus used to diagnose the results of this application, with a particular focus set on classification errors.

II. Methodology

A. Introduction of the Convolutional Autoencoder

A CAE is a deep neural network of the family of autoencoders. Its non-linear conception consists of an encoder part, which projects a time series onto a lower dimension space called the latent space, or the embedding space, using 1D-convolutions and fully connected (FC) layers. The second part of the CAE, the decoder, is tasked with reconstructing the input using this embedding representation through FC layers. The network is trained using a reconstruction task, calculated using a mean squared error between the original input and the reconstruction. Intuitively, the generated latent space is dense in information from the input and is thus used in various downstream applications, including unsupervised classification [4]. However, this latent space is uninterpretable, and these interpretability issues often block the analysis of the behavior of autoencoders in

Thomas Di Martino, Régis Guinvarc'h and Laetitia Thirion-Lefevre are with SONDRA Laboratory of CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, 91190, France (Email: thomas.di-martino@centralesupelec.fr; regis.guinvarch@centralesupelec.fr; laetitia.thirion@centralesupelec.fr).

Thomas Di Martino and Élise Colin are with ONERA, Département Traitement de l'Information et Systèmes, Université Paris-Saclay, Palaiseau, 91123, France (Email: thomas.di_martino@onera.fr; elise.colin@onera.fr).

the said applications. In supervised learning, gradient-based methods, including Grad-CAM, are used to tackle the difficulties of interpretation of the network. However, the original formulation of Grad-CAM [11] involves two pain points making its direct application to our context difficult:

- Firstly, its formulas are designed for image processing.
- Secondly, it is often applied in a classification context, on a vector of logits, at the network's output.

Thus, to adapt the Grad-CAM methodology to our context, we need to adjust its formulation to the processing of time series and the analysis of the embeddings layer of our autoencoder. To illustrate our contributions, we first present the state-of-the-art functioning of the Grad-CAM method and detail the changes made.

B. Introduction of Grad-CAM

As presented in [11], the goal of the Grad-CAM methodology is to "*obtain a class-discriminative localization map*". For ease of reading, we will use the same variables and function symbols as in [11]. The target localization map for a given class c is written $\mathbb{L}^c_{Grad-CAM}$ and is $h \times w$-dimensional, with $h, w \in \mathbb{Z}^{+*}$ the respective height and width of an input image. This map corresponds to a real-valued array, where the value of each pixel of coordinate (i, j) is equivalent to their relative importance in the decision to classify an input image as c. The map calculation is usually performed given a convolutional layer called A. Multiple convolutional layers can usually be found in a network. Thus, it is required to choose the layer onto which $\mathbb{L}_{Grad-CAM}$ is calculated. Another solution is calculating $\mathbb{L}$ for all convolutional layers and aggregating the maps into one.

To estimate $\mathbb{L}^c_{Grad-CAM}$, the original implementation of Grad-CAM proposes the following equation:

$$\mathbb{L}^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{1}$$

The components of this equation are the following:

- $ReLU(x) = max(x, 0)$.
- A^k is the feature map k of the convolutional layer A: it corresponds to the activation value of k-th filter of layer A applied to the input. Namely, the activation values are the output of an operation layer of a neural network, which serves, in turn as an input for the next layer.
- α^c_k is the gradient backpropagated from class prediction c to feature map k of the convolutional layer A.

α^c_k is originally defined as:

$$\alpha^c_k = \frac{1}{i * j} \sum_i \sum_j \frac{\delta y^c}{\delta A^k_{i,j}} \tag{2}$$

where $\frac{\delta y^c}{\delta A^k_{i,j}}$ is the partial derivative of the class c's logit y^c as a function of $A^k_{i,j}$, acting as the measure of the contribution of $A^k_{i,j}$ to the generation of y^c. A global average pooling of the backpropagated gradients is implemented to merge these contributions across the whole convolution layers. This calculation thus provides a measure of the contribution of the feature map k of the layer A to a prediction of class c. With the combination of α^c_k, representing the importance of the feature map k, and A^k representing the localization of the activations of the feature map k, we can generate a localized map of importance. Because of the use of activations, Grad-CAM relies on multi-level down-sampled importance information to build $\mathbb{L}^c_{Grad-CAM}$.

C. Adaptation of Grad-CAM

As mentioned before, the formulation of Grad-CAM does not allow for a direct translation to our use case, which is the application of CAEs to satellite image time series: we first need to take into account the sequential nature of our input data, being time series rather than images, but also the unsupervised and generative nature of the analyzed neural network.
In particular, instead of a prediction of a given class c, we now use the notation regarding the generation of an embedding dimension d. Also, instead of speaking of spatial coordinates of data points, we now mention temporal coordinates.

Thus, this leads us to redefine Eq. 1 and Eq. 2 as Eq. 3 and Eq. 4.

$$\mathbb{L}^d_{Grad-SLAM} = \sum_k \alpha^d_k A^k \tag{3}$$

Eq. 3 replaces Eq. 1 by removing the ReLU function, previously applied to the product of the gradient and activations. The original motivation behind the presence of the ReLU function was to retrieve only positive contributions to the gradient when focusing on a given class. Negative contributions are believed to be related to another class: indeed, in a classification context, in essence, the predictions of different classes are antagonistic, meaning that positive contributions for one will be negative contributions for others. It is not the case for the embedding dimensions. While we may look for a total decorrelation between the values of an embedding vector, we are still in a context where a negative contribution of a given data point to the

generation of a given embedding dimension may be as informative as a positive contribution.

$$\alpha_k^d = \frac{1}{t} \sum_t \frac{\delta e^d}{\delta A_t^k} \tag{4}$$

Eq. 4 replaces Eq. 2 by substituting the backpropagation of class-related gradient with the backpropagation of embeddings-related gradient, represented by the partial derivative of the embedding vector e. The choice to generate α values at the end of the encoder phase rather than at the end of the decoder is connected with our desire to enlighten the contributions of input dates to the generated embedding space. In an application relying on the output of the network and the associated reconstruction error rather than on the embedding space, one may find more utility in deriving contributions from the network's last layer. It may, for instance, show correlations between different dates within an input time series. We do not assess this potential in our work. To illustrate a concrete usage of the Grad-SLAM algorithm to diagnose the behavior of a CAE, we use the applicative context of unsupervised classification of SAR agricultural time series [4].

III. Application of Grad-SLAM to unsupervised classification of SAR time series

A. Unsupervised classification of SAR time series using CAE

In [4], a CAE was applied to retrieve crop type information, without supervision, from Sentinel-1 time series. As shown in Fig. 1, this application transforms input time series into 2D embedding vectors clustered using the k-Means algorithm. The resulting clusters are then assigned a class using a majority voting strategy to assess their quality and measure performance. While we show that unsupervised approaches are on par with conventional supervised techniques, no further explanation is given for the remaining classification errors.

Since the classification application relies on clustering the embedding space, the analysis of classification errors thus depends on interpreting the information content of that same embedding space. For that, the Grad-SLAM approach is an ideal tool. The study site consists of 61 Sentinel-1 acquisitions from 2017. The preprocessing and the metadata of the acquisitions are presented in detail in [4], [15].

The multitemporal Sentinel-1 images are labeled using 16 distinct agricultural classes, as shown in Fig. 2. Experimental design details can be found in the original study [4]. The detailed prediction results of our method on the test set are displayed in Fig. 3.

While these results are interesting as they display the potential to retrieve SAR time series class-level information without labels during training time, they lack the interpretation of the embedding space, which, once clustered, induces unsupervised classification errors.

B. Grad-SLAM for unsupervised classification explainability

The analysis of the results of Fig. 3 involves the usage of Grad-SLAM to diagnose the separation of classes in the embedding space by analyzing the date contribution to that space generation, as shown in Fig. 4. Within that space, three classes appear well separated: Cotton, Tomato, and Sugar Beet.

In Fig. 4a, we plot these three crop types' average VV intensity profiles, color-coded using their respective Grad-SLAM date importance profiles. Higher values of Grad-SLAM (dates colored in red) correspond to the most important dates for the generation of the embedding values of the given class, oppositely to lower values of Grad-SLAM (dates colored in blue). In the case of these three crop types, dates corresponding to crop transition periods (seeding, harvesting, tilling) are the most important to the model. This results in an efficient differentiation in the embedding space of classes where these periods induce drastic radiometric differences. However, these same periods do not allow for the distinction of all classes. When shifting the analysis towards classes that appear not well classified, such as Sweet Potato and Pepper, we obtain the results of Fig. 4b.

The Grad-SLAM illustration of the Cotton, Sweet Potato, and Pepper classes showcases that during the dates of importance to the CAE, the three crop types all undergo the same radiometric transitions, rendering them hard to separate in the embedding space. In addition, due to the class-majority voting strategy used to perform cluster-to-class assignation, Cotton's majority class outweighs the rest, resulting in Sweet Potato and Pepper crops being mistaken for Cotton crops, inducing low classification results for these crops. The modeling by the autoencoder of other periods of a crop's life, such as its growth peak, would ensure the differentiation of these classes. However, the highly restrictive 2-Dimensional embedding size limits the capacity of the current autoencoders. Thus, building on the results illustrated by the autoencoder, and the dates highlighted by Grad-SLAM, re-running the training of CAEs, with higher embedding dimensions could prevent class confusion between Cotton, Sweet Potato, and Pepper, with the additional dimensions focusing on the pre-senescence period of the crops, where the three classes appear separable.

Another category of crop type separation error lies in the classification results of the Fallow and Quinoa classes, with their predictions spread across many classes.

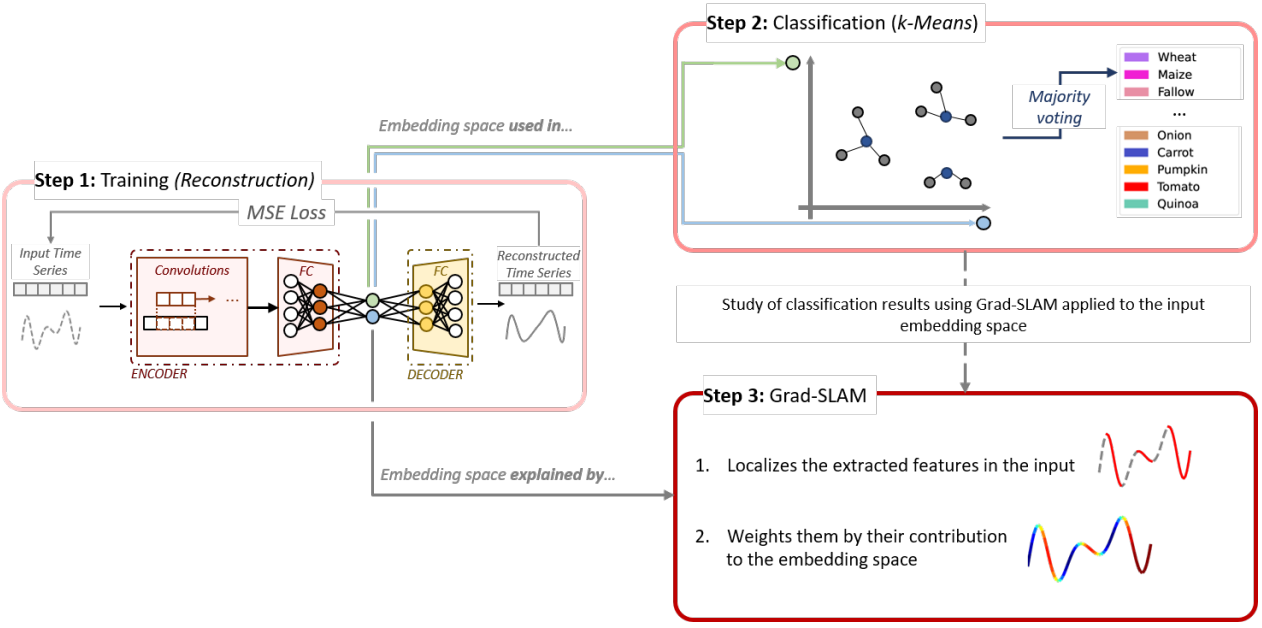

Fig. 1: Grad-SLAM application to unsupervised classification of time series.

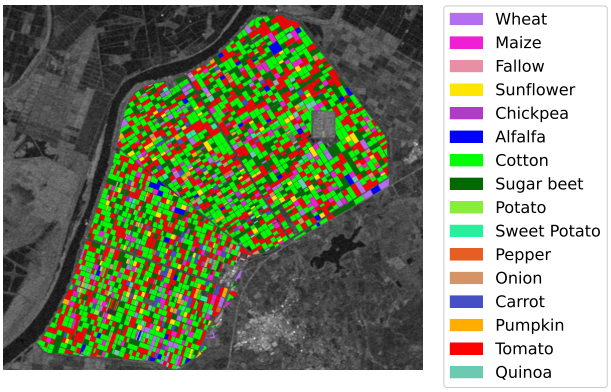

Fig. 2: Illustration of the BXII Sector and reference crop types data over a Sentinel 1 σ_0 VH polarization image dated January 3rd, 2017.

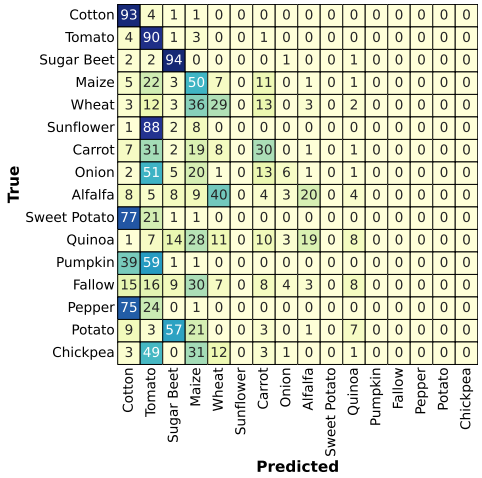

Fig. 3: Visualization of the CAE's classification performance over the test set [4] (in %, normalized by rows).

When analyzing their Grad-SLAM profile in Fig. 4c, it appears that the model focuses on dates irrelevant to the growth pattern of Quinoa, or the definition of a fallow crop parcel. Considering the lack of these two classes' radiometric transition periods, in opposition to the vast majority of the other classes, we assume that temporal features, characteristic of these two classes, are not modeled by the CAE, which results in their inaccurate and scattered latent representation and, in turn, erroneous unsupervised classification.

IV. Conclusion

In this work, we present an adaptation of the existing Grad-CAM methodology by introducing the Grad-SLAM method: it allows for the explainability of Convolutional Autoencoders applied to time series of satellite data. We detail the formulation changes between the original work and ours, induced by the switch from a supervised to an unsupervised paradigm and a switch from spatial to temporal information interpretation. We show that its application allows for a certain degree of explainability in the construction of the embedding space. In particular, it extracts the contribution of each input date to the generation of each specific embedding dimension. In turn, this embedding space was classified using a k-

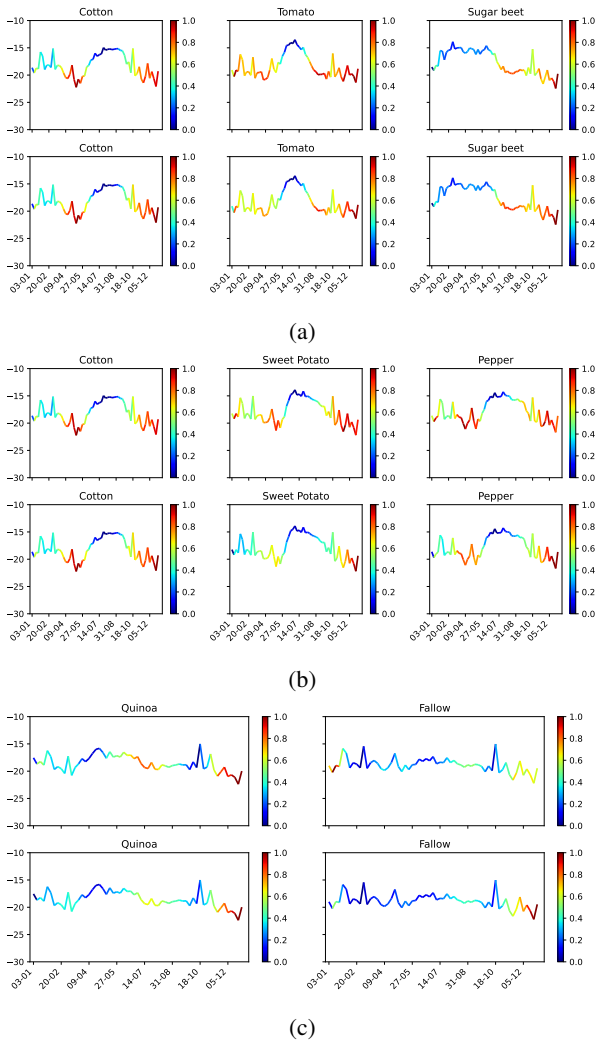

Fig. 4: Averaged VV intensity temporal profile, in dB superimposed with Normalized averaged Grad-SLAM date importance (higher is better). First row: per class $L^1_{Grad-SLAM}$; Second row: per class $L^2_{Grad-SLAM}$. (a) Well classified: Cotton, Tomato, and Sugar Beet (b) Class Confusion: Cotton, Sweet Potato, and Pepper (c) Absence of modelling: Fallow, and Quinoa.

Means algorithm, and class confusion arose from the classification. Using Grad-SLAM, these class confusions were explainable by highlighting the importance to the model of periods where the confused crop types are indistinguishable. While the Grad-SLAM algorithm was presented here in an unsupervised classification context, it also has potential for other CAE applications, such as anomaly detection, or inversion. Its extension to applications of autoencoding of multimodal time series, highlighting the contribution of various modalities, is also envisioned.

REFERENCES

[1] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sensing of Environment*, vol. 241, p. 111716, 2020.

[2] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.

[3] 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.

[4] T. Di Martino, R. Guinvarc'h, L. Thirion-Lefevre, and C. Koeniguer, "Beets or cotton? blind extraction of fine agricultural classes using a convolutional autoencoder applied to temporal sar signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[5] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.

[6] C. M. Gevaert, "Explainable ai for earth observation: A review including societal and regulatory perspectives," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102869, 2022.

[7] R. Roscher, B. Bohn, M. Duarte, and J. Garcke, "Explain it to me–facing remote sensing challenges in the bio-and geosciences with explainable machine learning," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 817–824, 2020.

[8] S. J. Newman and R. T. Furbank, "Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data," *Nature Plants*, vol. 7, no. 10, pp. 1354–1363, 2021.

[9] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102520, 2021.

[10] A. Wolanin, G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter, "Estimating and understanding crop yields with explainable deep learning in the indian wheat belt," *Environmental research letters*, vol. 15, no. 2, p. 024019, 2020.

[11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[12] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "Gee: A gradient-based explainable variational autoencoder for network anomaly detection," in *2019 IEEE Conference on Communications and Network Security (CNS)*, 2019, pp. 91–99.

[13] A. Bartler, D. Hinderer, and B. Yang, "Grad-lam: Visualization of deep neural networks for unsupervised learning," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1407–1411.

[14] L. Bergamasco, S. Saha, F. Bovolo, and L. Bruzzone, "An explainable convolutional autoencoder model for unsupervised change detection," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 1513–1519, 2020.

[15] A. Mestre-Quereda, J. M. Lopez-Sanchez, F. Vicente-Guijalba, A. W. Jacob, and M. E. Engdahl, "Time-Series of Sentinel-1 Interferometric Coherence and Backscatter for Crop-Type Mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4070–4084, 2020.