SELF ATTENTION DEEP GRAPH CNN CLASSIFICATION OF TIMES SERIES IMAGES FOR LAND COVER MONITORING

Ferdaous Chaabane, Safa Réjichi, Florence Tupin

SELF ATTENTION DEEP GRAPH CNN CLASSIFICATION OF TIMES SERIES IMAGES FOR LAND COVER MONITORING

Ferdaous Chaabane [(1)], Safa Réjichi [(1)], Florence Tupin [(2)]

(1) COSIM laboratory, SUP'COM, Carthage University, Tunisia
(2) Department of Image and Signal Processing, Telecom ParisTech, France

ABSTRACT

Time Series of Satellite Imagery (SITS) acquired by recent Earth observation systems represent an important source of information that supports several remote sensing applications related to monitoring the dynamics of the Earth's surface over large areas. A major challenge then is to design new deep learning models that can take into account intelligently the complementarity between temporal and spatial contexts that characterize these data structures. In this work, we propose to use an adapted self-attention convolutional neural network for spatio-temporal graphs classification that exploits both spatial and temporal dimensions. The graphs will be generated from a series of temporal images that are segmented into different regions. Those graphs are then classified using the Self-Attention Deep Graph CNN (DGCNN) model to highlight the temporal evolution of land cover areas through the construction of a spatio-temporal Map.

Index Terms— Graph based CNN, self-attention mechanism, satellite image time series, land cover classification, SOTAG (Spatial-Object Temporal Adjacency Graphs), etc.

1. INTRODUCTION

Given the spatial, structural and composition diversity of land cover, high spatial and temporal resolutions data are needed for earth ground dynamics analysis and monitoring. Space observation techniques have advanced significantly in recent years and the availability of time series of images with high spatial and temporal resolution, through the multiplication of performant sensors on the one hand, and high satellite revisit capacity on the other hand, opens up new perspectives for the study of land cover at fine scales. Thus, the availability of time series of satellite images made today possible to envisage the precise identification and monitoring of spatio-temporal regions and raises the question of the methodological approach to be adopted to process them.

Our work is part of this context. The general objective of our proposal is the exploitation of time series of segmented satellite images for analysis and monitoring of spatio-temporal entities (object-oriented) and the identification and highlighting of evolution patterns. SITS analysis work, already present in the literature [1][2], deals with the detection of changes by comparing series of segmented satellite images. The main problem that emerges is the alignment of objects in a time series. Indeed, when it comes to aligning pixels, it is enough to superimpose the images Satellites. It is more tedious for aligning objects because there is no one-by-one match between images. We already proposed in a previous work [3] an evolution Graph Based method which models temporal evolution of each SITS's region by a graph. Afterwards, a graph kernel based SVM classification is used to extract regions with similar temporal behaviors. Moreover, we proposed in [4] an original expert knowledge-based SITS analysis technique for land-cover monitoring and region dynamics assessing.

More recently, in [5], the authors examined and compared the performances of the RF, KNN, and SVM classifiers for land cover classification using Sentinel-2 image data. Furthermore, deep learning techniques have been recently adopted in the context of SITS data classification [6][7]. Authors in [6] performed satellite image time series classification using CNN on the temporal dimensions. In [7], the ability of RNNs, in particular, the LSTM model, to perform land cover classification considering SITS is evaluated. Recently, proposed a proposes a comparison between an SVM based spatio-temporal classification approach and the LSTM model for SITS analysis [8].

In this context, this paper proposes an original self attentive spatial temporal Deep Graph CNN based methodology which has three steps: (i) the detection of spatio-temporal entities (reference objects), (ii) the construction of evolution graphs and (iii) the Deep Graph CNN based classification of evolution graphs which makes possible to organize and highlight objects that evolve in the same way but also evolution patterns. We have adopted the same graph representation presented in our previous work [3][4] and we are proposing a new deep learning CNN based approach by introducing a self-attention mechanism that exploits both spatial and temporal dimension of SITS.

The paper is organized as follows. The first section presents the main steps of the graph regions construction. Then, the Self-attention CNN model is detailed in the third section. Finally, simulated and real data description and

experimental results are highlighted and discussed in the fourth section.

2. SPATIAL-OBJECT TEMPORAL ADJACENCY GRAPHS CONSTRUCTION

First, all SITS images are carefully pretreated (pansharpening and co-registration) to avoid misclassification results. Then, we identify the different regions included in the SITS by applying segmentation on each image [3].

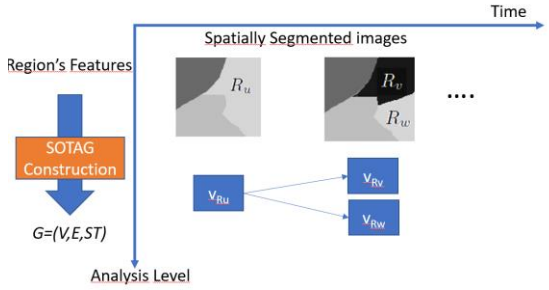

Figure 1. SOTAGs construction.

Spatially homogenous regions obtained after segmentation in each SITS image are temporally analyzed through regions' graph construction using the chronological order. To extract the temporal evolution of each region of SITS images, we consider successive couples of images.

A spatial-object temporal adjacency graph (SOTAG), noted G, is a graph representing temporal evolution of a region of SITS image such as $G = (V,E,ST)$ where V is the set of nodes (regions), E is the set of edges and ST is a function that, given a region, returns the corresponding multi-variate time series information : $ST(v_i) \in R^{T \times D}$ where T is the number of timestamps of the time series and D is the number of features on which the multi-variate time series is defined on. The D features are computed directly from image i describing region characteristics. They are clustered into three families: spectral, textural and spatial features [].

The set of edges E is derived from the set of V considering spatial adjacency. More in detail, for each $v_i, v_j \in V$ that are spatially adjacent (region v_i spatially touch region v_j). Finally, we define $N(v_i)$ as the set of neighborhood regions of a node v_i, where $N(v_i) = \{v_j | \exists (v_i, v_j) \in E\}$ and $|N(v_i)|$ is the cardinality of such a set.

3. SELF-ATTENTION DEEP GRAPH CNN BASED APPROACH

Figure 2 shows the main steps of the Self-Attention DGCNN approach. As illustrated, we first need to identify the spatio-temporal regions included in the SITS by applying segmentation on each image. Then, spatially homogenous regions are temporally analyzed through graph construction. Indeed, a SOTAG is constructed for each region of the first image. The obtained SOTAGs are at that point used as input to the Self-Attention module by considering both the spatial and the temporal dimensions. The resulted graphs are then aggregated to form spatio-temporal weighted graphs which are classified in order to determine the temporal evolution of each region of the first SITS image (stable, periodic evolution, abrupt change, etc.). a DGCNN is used for this purpose and outputs a spatio-temporal MAP.

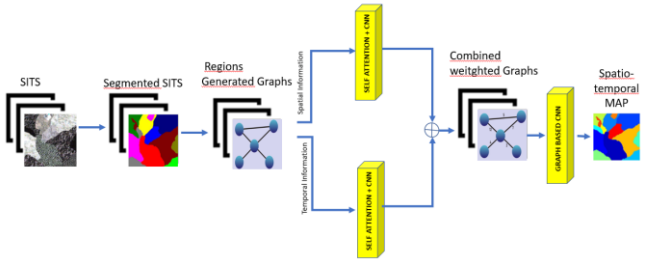

Figure 2. The Self-Attention Deep Graph CNN approach organizational chart

3.1. Self-Attention Mechanism

The Self-attention mechanism used in this work has been introduced in [8] It allows to consider both spatial nodes features and neighboring as well as the temporal graph topology.

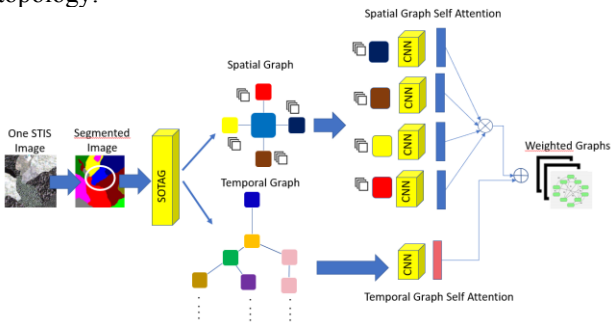

Figure 3. The Self-Attention Mechanism main steps.

Given a target region with its spatial neighborhood and temporal evolution (SOTAG) sets as input, The Self-Attention processes on one side the time series information and, on the other side the time series associated to the spatial neighborhood set. In both cases, the one dimensional convolutional neural network presented in [9] is employed as encoder network to extract the region embeddings. This encoder network operates on the temporal dimension of the SITS data explicitly modeling the sequential information it contains trough SOTAGs. For this encoder we use the one dimensional CNN as embedding extractor. Considering the neighborhood information, first the same one dimensional CNN model is applied over all the neighborhood regions. Then, since the two embeddings constitute complementary information that permits to characterize the sample to classify, they are successively combined together by means of a self-attention mechanism providing a new representation.

For the temporal region embedding, that we name as G^i_{temp}, the one dimensional CNN presented in [X3] is employed. As regards the neighborhood embedding, we remind that each target region v_i has an associated neighborhood set $N(v_i)$ with varying size. To aggregate together such varying-size information carried out by $N(v_i)$, we adopt a graph attention mechanism defined as follows:

$$G^i_{neigh} = |N(v_i)| \sum_{v_i \in N(v_i)} \alpha_{ij} G_{v_j} \quad (1)$$

where v_j is a region in the set $N(v_i)$, G_{vj} is the vector embedding of the region v_j. More precisely, the same one dimensional CNN model, based on the same set of learnable parameters, is employed over all the segments $v_j \in N(v_i)$ (cf. Figure 3). The attention coefficient α_{ij} that weights the contribution of the segment $v_j \in N(v_i)$ in the spatial neighborhood aggregation are obtained through the training process.

Once the target segment embedding (G^i_{temp}) and the spatial neighborhood embedding (G^i_{neigh}) are obtained, they are successively combined by means of a self-attention mechanism [9] with the goal of automatically weighting the contribution of the features extracted from the target segment as well as its spatial neighborhood. The output of this step is a representation which we refer to as G^i_{total}.

In the case of the combination of G^i_{temp} and G^i_{neigh}, the attention is not conditioned to any kind of information but it must only combine the target segment embedding and the neighborhood embedding together. To this end, we consider the attention mechanism originally introduced in []. Given $G^i_{total} = \{G^i_{temp}, G^i_{neigh}\}$, we attentively combine these two embeddings as follows:

$$G^i_{total} = \sum_{l \in \{temp,neigh\}} \alpha_l G^i_l \quad (2)$$

Where α_l with $l \in \{temp,neigh\}$ is defined as:

$$\alpha_l = \frac{\exp\left(v_a^T \tanh\left(W_a G^i_l + b_a\right)\right)}{\sum_{l' \in \{tem,neigh\}} \exp\left(\exp\left(v_a^T \tanh\left(W_a G^i_l + b_a\right)\right)\right)} \quad (3)$$

where matrix $W_a \in R^{d,d}$ and vectors $b_a, v_a \in R^d$ are parameters learned during the process. These parameters allow to combine G^i_{temp} and G^i_{neigh} (cf. Figure 3). The purpose of this procedure is to learn weights α_{temp} and α_{neigh}, and estimate the contribution of each of the embedding G^i_{temp} and G^i_{neigh}.

3.2. Supervised graph classification with Deep Graph CNN

The DGCNN architecture adopted in this work was proposed in [x1] using the graph convolutional layers from [x2] but with a modified Self-Attention Mechanism. [x1] introduces a Sort Pooling layer to generate an embedding representation for each given graph using as input the representations learned for each node via a stack of graph convolutional layers. For our case we use as an input the Self-Attention graphs G^i_{total} integrating the spatial and the temporal information. These graphs are then used as input to one-dimensional convolutional, max pooling, and dense layers that learn graph-level features suitable for predicting graph temporal labels (cf. Figure 4).

After the temporal behavior classification of each region of the first image of the SITS,

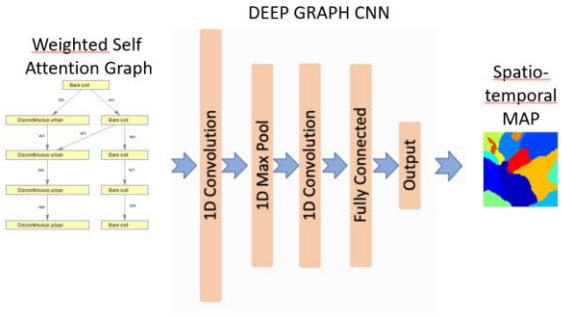

Figure 4. Deep Graph CNN main layers.

4. EXPERIMENTAL RESULTS

The proposed Self-Attention DGCNN approach is applied on both simulated and real SITS. Simulated ones are mainly used for validation (cf. Figure 5). Then, VHR-SITS covering Zaghouan region, located in the northern half of Tunisia (cf. Figure 6), is used to highlight the approach performance in a real context. This region is characterized by its ecosystem diversity.

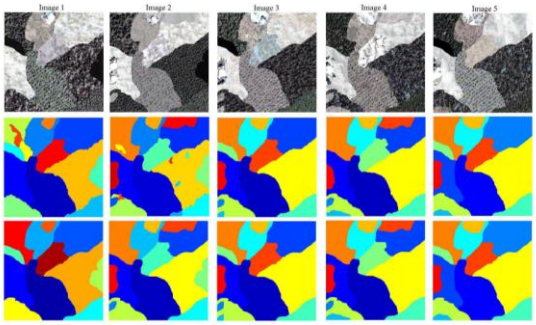

Figure 5. Simulated SITS images (first row), their segmentation results (second row), and their ground truths (third row).

Five images (four QUICKBIRD and one IKONOS images) taken at different dates were used for experimentations with pansharpening and coregistration as preprocessing. We used an SVM-based algorithm to segment SITS images. Besides, as ground truth map is available for synthesized SITS, evaluation is only done for simulated data. For real SITS, only a qualitative evaluation is presented. Due to lack of time we are going to present for this first abstract version the results obtained for the simulated data. Results concerning the real SITS will be added in the final version.

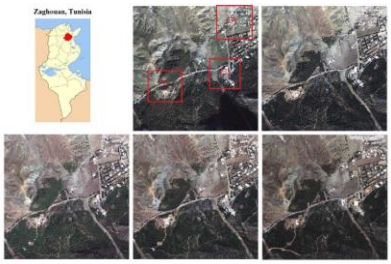

Figure6. Zaghouan region SITS and studied scenes (first row from left: February 6, 2001 and April 20, 2002, second row from left: September 11, 2004; June 21, 2005; and June 27, 2010).

For experimental evaluation of the final classification step, given a SITS region which temporal evolution is formalized by a SOTAG, the proposed method tries to find the temporal evolution (stable, periodic disappearance, abrupt appearance, etc.; evolution scenarios existing in the simulated SITS are illustrated in Figure 7) basing on the resulted Self-Attention Graph. The ground truth has been used to achieve quantitative evaluation and represents the right temporal evolutions that should be selected.

The use of the Self-Attention DGCNN allows the classification of almost all scenarios with 4 missed regions over 14 ones. This results in an overall accuracy equal to 82.23% and a kappa index equal to 0.7960. However, the simple DGCNN does not recognize 7 regions over 14 ones, which gives an overall accuracy equal to 79.02% and a kappa index equal to 0.705. Therefore, the addition of the Self-Attention module achieves better results. The performance of the Self-Attention based method may be explained by its ability to transform all the graph information into a new weighted spatio-temoporal space where the comparison between graphs and their classification is easier.

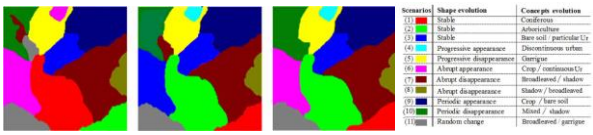

Figure 7. Simulated SITS spatio-temporal classification results : DGCNN, Self-Attention DGCNN and ground truth.

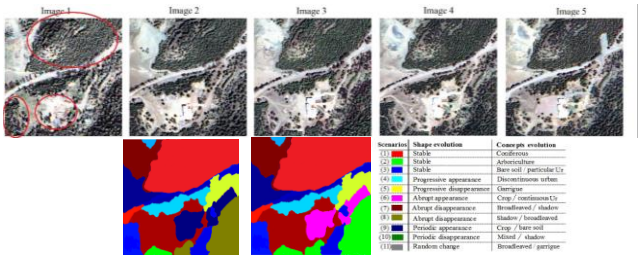

Figure 8. Real SITS spatio-temporal classification results : DGCNN (Left) and Self-Attention DGCNN (right).

Figure 8 shows a small area extracted from a real SITS which has both stable and changing regions (red circles highlight significant regions).

As we don't have the ground truth, we only perform a qualitative evaluation. We can see in Figure 8 that the Self-Attention DGCNN succeeds to classify correctly most of the stable regions and particularly the urban appearance region (new construction). However, the DGCNN misclassified several regions.

5. CONCLUSION

This works addressed the temporal classification of VHR SITS using A Self-Attention Deep Graph CNN approach which focuses on a global spatial evolution (region based). The qualitative results obtained for simulated data shows that the Self-Attention Graph mechanism which take into account the temporal and spatial dimensions improve the classification results comparing to a simple Deep Graph CNN technique. However, the added Self-Attention module is consuming and need many samples for training. To conclude, we may suggest the choice of the appropriate technique according to the nature of regions and the consumption time.

6. REFERENCES

[1] A. Julea, N. Méger, P. Bolon, C. Rigotti, M.P. Doin, C. Lasserre, E. Trouvé, V. Lazarescu, "Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns," *IEEE TGRS*, vol. 49, Issue 4, pp. 1417-1430, 2011.

[2] C. Kurtz, F. Petitjean, P. Gancarski, "A context-based approach for the classification of Satellite Image Time Series," *IEEE IGARSS*, pp. 495-498, Vancouver, Canada, 24-29 July 2011.

[3] S. Rejichi and F. Chaabane, "Feature extraction using PCA for VHR satellite image time series spatio-temporal classification," *IEEE IGARSS*, pp.485-488, Milano, Italy, 26-31 July 2015.

[4] S. Rejichi, F. Chaabane, F. Tupin, "Expert knowledge-based method for Satellite Image Time Series analysis and interpretation," *IEEE (J-STARS)*, vol. 8, no. 5, pp. 2138-2150, May 2015.

[5] P. T. Noi, M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery," *Sensors* 2018.

[6] C. Pelletier, G. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," Remote Sens., vol. 11, no. 5, p. 523, Mar. 2019

[7] D. Ienco, R. Gaetano, C. Dupaquier, "Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks," IEEE Geoscience and Remote Sensing Letters, (14)10, October 2017.

[8] A. M. Censi et al, "Attentive Spatial Temporal Graph CNN for LC Mapping From Multi Temporal Remote Sensing Data, MDL4EO: Machine and Deep Learning for Earth Observation data, " DOI: 10.1109/ACCESS.2021.3055554, January 2021.

[9] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, and R. Interdonato, "Weakly supervised learning for land cover mapping of satellite image time series via attention-based CNN," IEEE Access, 8:179547–179560, 2020.

[10] M. Zhang, Z. Cui, M. Neumann, Y. Chen, "An End-to-End Deep Learning Architecture for Graph Classification,"AAAI-2018.

[19] T. N. Kipf and M. Welling, "Semi-supervised Classification with Graph Convolutional Networks," ICLR 2017.