DEEP CLUSTERING METHODS STUDY APPLIED TO SATELLITE IMAGES TIME SERIES

Baptiste Lafabregue[1], *Anne Puissant*[2], *Jonathan Weber*[3], *Germain Forestier*[3]

[1]ICube, Université de Strasbourg, France, lafabregue@unistra.fr,
[2]LIVE, Université de Strasbourg, France, anne.puissant@live-cnrs.unistra.fr,
[3]IRIMAS, Université de Haute Alsace, France, {germain.forestier,jonathan.weber}@uha.fr

ABSTRACT

Clustering is an essential tool for data analysis and visualization. It is particularly useful in case of a lack of labels, which prevent the use of supervised methods. The analysis of satellite images is particularly prone to this problem, especially when studied as time series, because the access to this type of data is still recent. Among all clustering methods, the ones based on Deep Neural Networks (DNNs) have seen an increasing interest lately, but only a few works have been conducted on time series yet. This paper aims to give more insight on how current clustering methods based on DNNs can be applied to Satellite Images Time Series (SITS) and it shows that with a proper configuration they can perform better compared to classical non-deep methods.

***Index Terms*—** Image time-series, Clustering, Deep learning, remote sensing

1. INTRODUCTION

The lack of data is not a major problem anymore when studying Satellite Images Time Series (SITS). Indeed, the introduction of a new satellite constellation can now provide regular and readily available earth observations. These satellites provide useful data that can be used in many applications, from the management of natural disasters or urban planning. However, labeling these time series remains a difficult and time-consuming task that often prevents the use of supervised methods. In this context, clustering methods offer a useful tool to assist the user during the data analysis. Nonetheless, most of these clustering methods rely on a norm-based distance function that implies a fixed mapping between points in two time-series and are therefore a sensitivity to noise and misalignment in time [1]. Multiple approaches have been proposed in the literature to answer this problem and are mostly based either on a specific metric, such as Dynamic Time Warping (DTW) [2], or the application of data transformations to reduce or remove the time dimension, such as Symbolic Aggregate ApproXimation (SAX) [3].

This work was part of a PhD supported by the ANR TIMES project (ANR-17- CE23-0015) of the National French Agency.

Deep Clustering methods can be classified among the *transformation* group as they consist in learning a data representation. These methods have attracted increasing attention in the past few years, mostly in the image processing domain. Recently, some works have been conducted to adapt and evaluate these methods for time series analysis [4, 5]. The obtained results show a high potential to improve the state of the art, but with mixed results partially depending on the analyzed dataset.

In this paper, we study a set of deep clustering approaches on two SITS datasets to evaluate what types of configuration are more fitted to this kind of data, and if they are relevant when compared to non-deep classical approaches.

2. BACKGROUND

2.1. Deep Learning and clustering

Most clustering methods that involve Deep Neural Networks (DNNs) consist in training a network to learn a representation that will then be fed to a classical clustering algorithm, usually the K-Means algorithm. Hence, given a dataset X, the deep clustering task can be viewed as partitioning the set Z obtained with the non-linear mapping function $f_\Theta : X \to Z$, where f_Θ is a DNN called an encoder, and Θ are its learnable parameters. Z is often called the latent space, or latent representation of X. Therefore, we aim to find a proper way to obtain a function f_Θ that generates a relevant and easy to partition representation. To do so, multiple elements have to be taken into account.

2.2. DNNs decomposition

In [4], the authors proposed to decompose the DNNs into three elements:

Encoder architecture: the term architecture refers, in this article, to the set of layers and their hyperparameters. We can find four major families of layers applied to time series in the literature: Fully connected layers, Convolutional layers, Recurrent layers, and Attention layers.

Pretext loss: this loss is used to train the encoder to extract relevant features from the initial data. In an unsupervised

 DOI: https://doi.org/10.1109/IGARSS46834.2022.9884322

setting, we cannot use labels to directly train the DNN for class prediction. Therefore, we need to find a *pretext/proxy* task, assuming that the features learned by the encoder to solve this task will also be useful for the clustering task. Multiple pretext losses exist in the literature, such as the reconstruction loss for autoencoders [6] or its variation for the denoising autoencoders, or the triplet loss proposed for times series in [7].

Clustering loss: this loss is used to train the encoder to project the data into a latent space that is easier to partition. As the pretext loss is not designed for classification, the resulting projection may be either too sparse or with intricate clusters. This loss usually comes in complement to a pretext loss. It is used after the pretext loss either by replacing it or in parallel. These losses mainly focus on creating more dense clusters in the latent space [8].

3. EXPERIMENTS

3.1. Evaluated methods

There is a very large number deep clustering methods in the literature. Hence, we cannot evaluate all of them. Consequently, the choice was made to only evaluate the methods presented in the study from [4] as they cover a large spectrum of methods. These methods are constructed as a set of combinations among different architectures, pretext losses, and clustering losses. But, in this paper, as the number of methods remains high, we only displayed the most performing combinations selected by average rank over the datasets.

A total of 226 combinations have been evaluated. The combinations involve:

8 architectures : a multilayer perceptron (*MLP*) - three bidirectional Recurrent Neural Networks (RNN), one with LSTM units (*BLSTM*), one with GRU units (*BGRU*) and one with vanilla RNN units (*BRNN*) - an encoder based on dilated RNN, an encoder based on dilated Convolutional Neural Networks (CNN) (*DCNN*) - a CNN with only standard CNN layers followed by a Average Pooling layer (*SCNN*) - a CNN with three residual blocks (*RCNN*).

8 pretext losses : the classical reconstruction loss (*rec*) - the ELBO loss for Variational AutoEncoders (*vae*) - the triplet loss [7] with K equals to 1,2,5, 10 and combined (*tripletKxxx*)- a joint reconstruction loss [9] (*multi_rec*).

5 clustering losses : the loss from Deep Embedded Clustering method [8] (*DEC*) - the loss from Improved DEC method [10] - the loss from Deep Embedded Regularized Clustering [9] (*DEPICT*) - an adaptation of VAE for clustering called Variational Deep Embedding [11] (*VADE*) - an adaptation of Generative Adversarial Networks for clustering called ClusterGAN [12] - no use of clustering loss (*None*).

Each combination is named as follow : <*architecure*>-<*pretext loss*>-<*clustering loss*>. A more in-depth and detailed review is available in [4]. For the deep clustering methods, we used the code provided by the authors of [4].

We also used for comparison three non-deep methods based on the K-Means algorithm. The first one is the classical version based on Euclidean distance and arithmetic mean. The second one is using DTW to compute the dissimilarity and DBA (DTW Barycenter Averaging) [13] for the mean. The last one is the KShape method [14].

3.2. Data and evaluation procedure

Analysis land cover dynamics, and more particularly agriculture monitoring, is a major topic that has seen a lot of applications involving temporal data [15, 16]. For this experiment, we have selected two datasets.

The first one consists of four crop classes, traditional orchards, intensive orchards, standard meadows, and wet meadows located around Strasbourg (North-East France). The data is composed of 39 Sentinel-2 multispectral images on four bands (Red, Green, Blue, NIR) captured between 2017 and 2018 where at most 50% of annotated data are covered by clouds (the cloudy time steps are filled by linear interpolation). Some examples of the different classes are presented in Fig. 1. The reference data comes from expert annotations labeled by visual interpretation[1]. This dataset is actually divided into three sub-datasets, one per task. One task is to discriminate between the two orchards classes (24937 time series for the test set and 22064 for train). Another is to discriminate between the two meadows classes (241293 for test and 23445 for train). The last one is to discriminate between orchards and meadows (combination of the two others).

The second one consists of 12 crop classes located near Toulouse (Southwest France). The data is composed of 11 Formosat-2 multispectral images on three bands (Red, Green, NIR) without clouds captured in 2007 [2]. Some examples of two crop classes are presented in Fig. 2. The reference data is extracted from the farmer's declaration to the EEA's Common Agricultural Policy. A total of 1974 time series compose the test set and 9869 for the train.

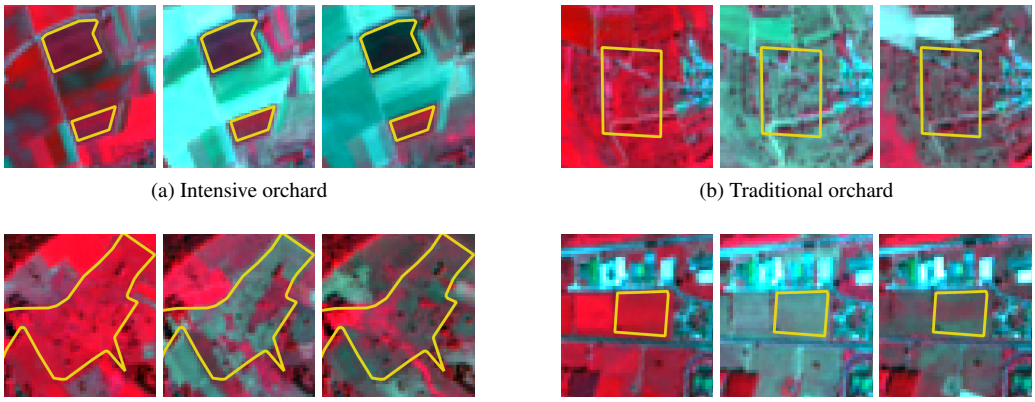
For both datasets, the train/test split is based on polygons (time series from the same polygon are only in one set) but evaluated on pixel time series cluster assignment. The train set is used to learn the DNNs models' parameters. Both datasets are z-normalized as a pre-processing. Performance is measured by the average Normalized Mutual Information (NMI) on five repetitions.

4. RESULTS

The results for each dataset are presented in Table 1. For the deep clustering combinations, we have only displayed the

[1]Provided by the *LIVE, Unité Mixte de Recherche CNRS-Unistra*, Strasbourg, France, available at https://www.kaggle.com/baptistel/meadows-vs-orchards

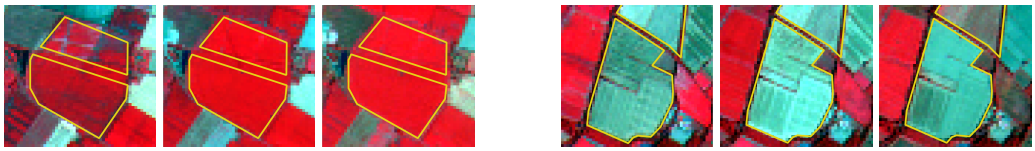
[2]Provided by the *Centre d'Études Spatiales de la Biosphère (CESBIO) Unité Mixte de Recherche CNES-CNRS-IRD-UPS*, Toulouse, France.



(a) Intensive orchard (b) Traditional orchard

(c) Standard meadow (d) Wet meadow

Fig. 1. Image crops from the *Orchard/Meadow* dataset from 08/05/2018, 02/07/2018, and 16/08/2018

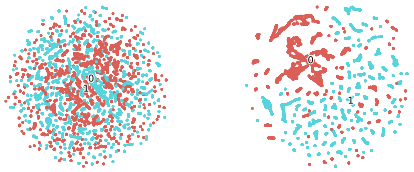


(a) Irrigated corn (b) Wheat

Fig. 2. Image crops from the *Crops* dataset from 01/07/2007, 11/08/2007, and 15/09/2007

top 5 combinations based on their average rank on the four datasets' train sets but full results are available online [3]. This five deep clustering methods obtain a higher average rank than the three non-deep methods. One could notice that the three first datasets on orchards and meadows can't be considered as independent. However, on one hand, the combination *DCNN-rec-DEPICT* obtains higher results on all datasets than all non-deep methods and manage to obtain a more separable data space as illustrated by the UMAP [17] visualization in Fig. 3. On the other hand, the other combinations result in more mitigated performances. Moreover, most combinations lead to worst performances than the Euclidean K-Means algorithm. This highlights the difficulty of choosing the most fitted hyperparameters for these methods as we don't have any labels to help select them for each dataset.

The results partially validate the first observations made in [4]. Indeed, they showed that both CNN-based architectures and reconstruction-based pretext losses gave better than other approaches. However, on the clustering loss, it was shown that not using one gives better results. In our study, the clustering losses from the clustering method *DEPICT* and *IDEC* improve slightly the performances for the combination *DCNN-rec*, but on average on all combinations, it actually degrades slightly the performances. For example, the combination *SCNN-rec* obtains better performances without any clustering loss (*None*) while obtaining the best results on both discriminating between meadows classes and between orchards and meadows.



(a) Raw data (b) *DCNN-rec-DEPICT*

Fig. 3. Visualization with UMAP method of the dataset Orchards with the raw data and the latent space computed by the combination *DCNN-rec-DEPICT*. Blue is for traditional orchards and red for intensive orchards.

5. CONCLUSION

Our study shows that Deep clustering methods, when well parameterized can outperform classical non-deep methods on satellite images time series. Even though these methods show a lot of potentials, we also show that the choice of configuration and hyperparameters can have a strong influence on

[3] https://github.com/blafabregue/TimeSeriesDeepClustering/blob/main/paper_results/igarss2022_results.csv

Table 1. ARI average results for the top 5 deep combination (by rank on train set) and non-deep methods. The best performance for each dataset is highlighted in bold. The average rank is one over all 229 methods (226 deep combinations + 3 non-deep methods) evaluated.

Dataset	*DCNN-rec-None*	*DCNN-rec-DEPICT*	*DCNN-rec-IDEC*	*DCNN-multi_rec-DEPICT*	*SCNN-rec-None*	*Euclidean K-Means*	*DTW K-Means*	*Kshape*
Orchards vs Meadows	0.155	0.136	0.155	0.108	**0.193**	0.109	0.057	0.062
Meadows	0.127	0.128	0.152	0.158	**0.200**	0.113	0.030	0.000
Orchards	0.636	**0.668**	0.639	0.609	0.564	0.622	0.278	0.029
Crops	0.507	**0.752**	0.497	0.683	0.490	0.557	0.429	0.531
Average rank	17.75	**11.5**	16	19	17.25	23	74.75	87.5

the performance of the method. Therefore, a further study should be conducted to investigate the potential correlation between some specific parameters and the datasets' characteristics on the quality of the learned representation for the clustering task.

6. REFERENCES

[1] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," *Data Min. Know. Disc.*, vol. 7, no. 4, pp. 349–371, 2003.

[2] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *International Congress on Acoustics*, 1971, pp. 65–69.

[3] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Min. Know. Disc.*, vol. 15, pp. 107–144, 2007.

[4] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier, "End-to-end deep representation learning for time series clustering: a comparative study," *Data Min. Know. Disc.*, pp. 1–53, 2021.

[5] N. Tavakoli, S. Siami-Namini, M.A. Khanghah, F.M.. Soltani, and A.S. Namin, "An autoencoder-based deep learning approach for clustering time series data," *SN Applied Sciences*, vol. 2, no. 5, pp. 1–25, 2020.

[6] D.H. Ballard, "Modular learning in neural networks.," in *AAAI*, 1987, pp. 279–284.

[7] J. Franceschi, A. Dieuleveut, and M. Jaggi, "Unsupervised scalable representation learning for multivariate time series," in *NeurIPS*, 2019, pp. 4652–4663.

[8] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *ICML*, 2016, pp. 478–487.

[9] K.G̃hasedi Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proceedings of the ICCV*, 2017, pp. 5736–5745.

[10] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation.," in *IJCAI*, 2017, pp. 1753–1759.

[11] X. Li, Z. Chen, L.K.M. Poon, and N.L. Zhang, "Learning latent superstructures in variational autoencoders for deep multidimensional clustering," in *ICLR*, 2018.

[12] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in *AAAI*, 2019, vol. 33, pp. 4610–4617.

[13] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011.

[14] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD*, 2015, pp. 1855–1870.

[15] T. Lampert, B. Lafabregue, and P. Gançarski, "Constrained distance based k-means clustering for satellite image time-series," in *IGARSS 2019-2019 IGARSS*. IEEE, 2019, pp. 2419–2422.

[16] C. Marais Sicre, J. Inglada, R. Fieuzal, F. Baup, S. Valero, J. Cros, M. Huc, and V. Demarez, "Early detection of summer crops using high spatial resolution optical image time series," *Remote Sensing*, vol. 8, no. 7, pp. 591, 2016.

[17] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *Journal of Open Source Software*, vol. 3, no. 29, 2018.